Low-density graph codes that are optimal for source/channel coding and binning

Martin J. Wainwright
Dept. of Statistics, and
Dept. of Electrical Engineering and Computer Sciences
University of California, Berkeley
wainwrig@{eecs,stat}.berkeley.edu

Emin Martinian
Tilda Consulting, Inc.
Arlington, MA
emin@alum.mit.edu

Technical Report 730,
Department of Statistics, UC Berkeley,
April 2007

Abstract

We describe and analyze the joint source/channel coding properties of a class of sparse graphical codes based on compounding a low-density generator matrix (LDGM) code with a low-density parity check (LDPC) code. Our first pair of theorems establish that there exist codes from this ensemble, with all degrees remaining bounded independently of block length, that are simultaneously optimal as both source and channel codes when encoding and decoding are performed optimally. More precisely, in the context of lossy compression, we prove that finite degree constructions can achieve any pair (R, D) on the rate-distortion curve of the binary symmetric source. In the context of channel coding, we prove that finite degree codes can achieve any pair (C, p) on the capacity-noise curve of the binary symmetric channel. Next, we show that our compound construction has a nested structure that can be exploited to achieve the Wyner-Ziv bound for source coding with side information (SCSI), as well as the Gelfand-Pinsker bound for channel coding with side information (CCSI). Although the current results are based on optimal encoding and decoding, the proposed graphical codes have sparse structure and high girth that renders them well-suited to message-passing and other efficient decoding procedures.

Keywords: Graphical codes; low-density parity check code (LDPC); low-density generator matrix code (LDGM); weight enumerator; source coding; channel coding; Wyner-Ziv problem; Gelfand-Pinsker problem; coding with side information; information embedding; distributed source coding.

1 Introduction

Over the past decade, codes based on graphical constructions, including turbo codes [3] and low-density parity check (LDPC) codes [17], have proven extremely successful for channel coding problems. The sparse graphical nature of these codes makes them very well-suited to decoding using efficient message-passing algorithms, such as the sum-product and max-product algorithms. The asymptotic behavior of iterative decoding on graphs with high girth is well-characterized by the density evolution method [25, 39], which yields a useful design principle for choosing degree distributions. Overall, suitably designed LDPC codes yield excellent practical performance under iterative message-passing, frequently very close to Shannon limits [7].

However, many other communication problems involve aspects of lossy source coding, either alone or in conjunction with channel coding, the latter case corresponding to joint source-channel coding problems. Well-known examples include lossy source coding with side information (one variant corresponding to the Wyner-Ziv problem [45]), and channel coding with side information (one variant being the Gelfand-Pinsker problem [19]). The information-theoretic schemes achieving the optimal rates for coding with side information involve delicate combinations of source and channel coding. For problems of this nature—in contrast to the case of pure channel coding—the use of sparse graphical codes and message-passing algorithm is not nearly as well understood. With this perspective in mind, the focus of this paper is the design and analysis sparse graphical codes for lossy source coding, as well as joint source/channel coding problems. Our main contribution is to exhibit classes of graphical codes, with all degrees remaining bounded independently of the blocklength, that simultaneously achieve the information-theoretic bounds for both source and channel coding under optimal encoding and decoding.

1.1 Previous and ongoing work

A variety of code architectures have been suggested for lossy compression and related problems in source/channel coding. One standard approach to lossy compression is via trellis-code quantization (TCQ) [26]. The advantage of trellis constructions is that exact encoding and decoding can be performed using the max-product or Viterbi algorithm [24], with complexity that grows linearly in the trellis length but exponentially in the constraint length. Various researchers have exploited trellis-based codes both for single-source and distributed compression [6, 23, 37, 46] as well as information embedding problems [5, 15, 42]. One limitation of trellis-based approaches is the fact that saturating rate-distortion bounds requires increasing the trellis constraint length [43], which incurs exponential complexity (even for the max-product or sum-product message-passing algorithms).

Other researchers have proposed and studied the use of low-density parity check (LDPC) codes and turbo codes, which have proven extremely successful for channel coding, in application to various types of compression problems. These techniques have proven particularly successful for *lossless* distributed compression, often known as the Slepian-Wolf problem [18, 40]. An attractive feature is that the source encoding step can be transformed to an equivalent noisy channel decoding problem, so that known constructions and iterative algorithms can be leveraged. For *lossy* compression, other work [31] shows that it is possible to approach the binary rate-distortion bound using LDPC-like codes, albeit with degrees that grow logarithmically with the blocklength.

A parallel line of work has studied the use of low-density generator matrix (LDGM) codes, which correspond to the duals of LDPC codes, for lossy compression problems [30, 44, 9, 35, 34]. Focusing on binary erasure quantization (a special compression problem dual to binary erasure channel coding), Martinian and Yedidia [30] proved that LDGM codes combined with modified message-

passing can saturate the associated rate-distortion bound. Various researchers have used techniques from statistical physics, including the cavity method and replica methods, to provide non-rigorous analyses of LDGM performance for lossy compression of binary sources [8, 9, 35, 34]. In the limit of zero-distortion, this analysis has been made rigorous in a sequence of papers [12, 32, 10, 14]. Moreover, our own recent work [28, 27] provides rigorous upper bounds on the effective rate-distortion function of various classes of LDGM codes. In terms of practical algorithms for lossy binary compression, researchers have explored variants of the sum-product algorithm [34] or survey propagation algorithms [8, 44] for quantizing binary sources.

1.2 Our contributions

Classical random coding arguments [11] show that random binary linear codes will achieve both channel capacity and rate-distortion bounds. The challenge addressed in this paper is the design and analysis of codes with *bounded graphical complexity*, meaning that all degrees in a factor graph representation of the code remain bounded independently of blocklength. Such sparsity is critical if there is any hope to leverage efficient message-passing algorithms for encoding and decoding. With this context, the primary contribution of this paper is the analysis of sparse graphical code ensembles in which a low-density generator matrix (LDGM) code is compounded with a low-density parity check (LDPC) code (see Fig. 2 for an illustration). Related compound constructions have been considered in previous work, but focusing exclusively on channel coding [16, 36, 41]. In contrast, this paper focuses on communication problems in which source coding plays an essential role, including lossy compression itself as well as joint source/channel coding problems. Indeed, the source coding analysis of the compound construction requires techniques fundamentally different from those used in channel coding analysis. We also note that the compound code illustrated in Fig. 2 can be applied to more general memoryless channels and sources; however, so as to bring the primary contribution into sharp focus, this paper focuses exclusively on binary sources and/or binary symmetric channels.

More specifically, our first pair of theorems establish that for any rate $R \in (0, 1)$, there exist codes from compound LDGM/LDPC ensembles with all degrees remaining bounded independently of the blocklength that achieve both the channel capacity and the rate-distortion bound. To the best of our knowledge, this is the first demonstration of code families with bounded graphical complexity that are simultaneously optimal for both source and channel coding. Building on these results, we demonstrate that codes from our ensemble have a naturally "nested" structure, in which good channel codes can be partitioned into a collection of good source codes, and vice versa. By exploiting this nested structure, we prove that codes from our ensembles can achieve the information-theoretic limits for the binary versions of both the problem of lossy source coding with side information (SCSI, known as the Wyner-Ziv problem [45]), and channel coding with side information (CCSI, known as the Gelfand-Pinsker [19] problem). Although these results are based

on optimal encoding and decoding, a code drawn randomly from our ensembles will, with high probability, have high girth and good expansion, and hence be well-suited to message-passing and other efficient decoding procedures.

The remainder of this paper is organized as follows. Section 2 contains basic background material and definitions for source and channel coding, and factor graph representations of binary linear codes. In Section 3, we define the ensembles of compound codes that are the primary focus of this paper, and state (without proof) our main results on their source and channel coding optimality. In Section 4, we leverage these results to show that our compound codes can achieve the information-theoretic limits for lossy source coding with side information (SCSI), and channel coding with side information (CCSI). Sections 5 and 6 are devoted to proofs that codes from the compound ensemble are optimal for lossy source coding (Section 5) and channel coding (Section 6) respectively. We conclude the paper with a discussion in Section 7. Portions of this work have previously appeared as conference papers [28, 29, 27].

2 Background

In this section, we provide relevant background material on source and channel coding, binary linear codes, as well as factor graph representations of such codes.

2.1 Source and channel coding

A binary linear code $\mathbb{C}$ of block length n consists of all binary strings $x \in \{0,1\}^n$ satisfying a set of $m < n$ equations in modulo two arithmetic. More precisely, given a parity check matrix $H \in \{0,1\}^{m \times n}$, the code is given by the null space

$$\mathbb{C} \quad := \quad \{x \in \{0,1\}^n \;\; | \;\; Hx = 0\}\,. \tag{1}$$

Assuming the parity check matrix H is full rank, the code $\mathbb{C}$ consists of $2^{n-m} = 2^{nR}$ codewords, where $R = 1 - \frac{m}{n}$ is the code rate.

Channel coding: In the channel coding problem, the transmitter chooses some codeword $x \in \mathbb{C}$ and transmits it over a noisy channel, so that the receiver observes a noise-corrupted version Y. The channel behavior is modeled by a conditional distribution $\mathbb{P}(y \mid x)$ that specifies, for each transmitted sequence Y, a probability distribution over possible received sequences $\{Y = y\}$. In many cases, the channel is memoryless, meaning that it acts on each bit of $\mathbb{C}$ in an independent manner, so that the channel model decomposes as $\mathbb{P}(y \mid x) = \prod_{i=1}^n f_i(x_i; y_i)$ Here each function $f_i(x_i; y_i) = \mathbb{P}(y_i \mid x_i)$ is simply the conditional probability of observing bit y_i given that x_i was transmitted. As a simple example, in the binary symmetric channel (BSC), the channel flips each

transmitted bit x_i with probability p, so that $\mathbb{P}(y_i \mid x_i) = (1-p)\,\mathbb{I}[x_i = y_i] + p\,(1 - \mathbb{I}[x_i \neq y_i])$, where $\mathbb{I}(A)$ represents an indicator function of the event A. With this set-up, the goal of the receiver is to solve the *channel decoding problem*: estimate the most likely transmitted codeword, given by $\widehat{x} := \arg\max_{x \in \mathbb{C}} \mathbb{P}(y \mid x)$. The Shannon capacity [11] of a channel specifies an upper bound on the rate R of any code for which transmission can be asymptotically error-free. Continuing with our example of the BSC with flip probability p, the capacity is given by $C = 1 - h(p)$, where $h(p) := -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.

Lossy source coding: In a lossy source coding problem, the encoder observes some source sequence $S \in \mathcal{S}$, corresponding to a realization of some random vector with i.i.d. elements $S_i \sim \mathbb{P}_S$. The idea is to compress the source by representing each source sequence S by some codeword $x \in \mathbb{C}$. As a particular example, one might be interested in compressing a *symmetric Bernoulli source*, consisting of binary strings $S \in \{0,1\}^n$, with each element S_i drawn in an independent and identically distributed (i.i.d.) manner from a Bernoulli distribution with parameter $p = \frac{1}{2}$. One could achieve a given compression rate $R = \frac{m}{n}$ by mapping each source sequence to some codeword $x \in \mathbb{C}$ from a code containing $2^m = 2^{nR}$ elements, say indexed by the binary sequences $z \in \{0,1\}^m$. In order to assess the quality of the compression, we define a source decoding map $x \mapsto \widehat{S}(x)$, which associates a source reconstruction $\widehat{S}(x)$ with each codeword $x \in \mathbb{C}$. Given some distortion metric $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$, the *source encoding problem* is to find the codeword with minimal distortion—namely, the optimal encoding $\widehat{x} := \arg\min_{x \in \mathbb{C}} d(\widehat{S}(x), S)$. Classical rate-distortion theory [11] specifies the optimal trade-offs between the compression rate R and the best achievable average distortion $D = \mathbb{E}[d(\widehat{S}, S)]$, where the expectation is taken over the random source sequences S. For instance, to follow up on the Bernoulli compression example, if we use the Hamming metric $d(\widehat{S}, S) = \frac{1}{n}\sum_{i=1}^n |\widehat{S}_i - S_i|$ as the distortion measure, then the rate-distortion function takes the form $R(D) = 1 - h(D)$, where h is the previously defined binary entropy function.

We now provide definitions of "good" source and channel codes that are useful for future reference.

Definition 1. (a) A code family is a *good D-distortion binary symmetric source code* if for any $\epsilon > 0$, there exists a code with rate $\mathrm{R} < 1 - h(D) + \epsilon$ that achieves Hamming distortion less than or equal to D.
(b) A code family is a *good BSC(p)-noise channel code* if for any $\epsilon > 0$ there exists a code with rate $\mathrm{R} > 1 - h(p) - \epsilon$ with error probability less than ϵ.

2.2 Factor graphs and graphical codes

Both the channel decoding and source encoding problems, if viewed naively, require searching over an exponentially large codebook (since $|\mathbb{C}| = 2^{nR}$ for a code of rate R). Therefore, any practically

useful code must have special structure that facilitates decoding and encoding operations. The success of a large subclass of modern codes in use today, especially low-density parity check (LDPC) codes [17, 38], is based on the sparsity of their associated factor graphs.

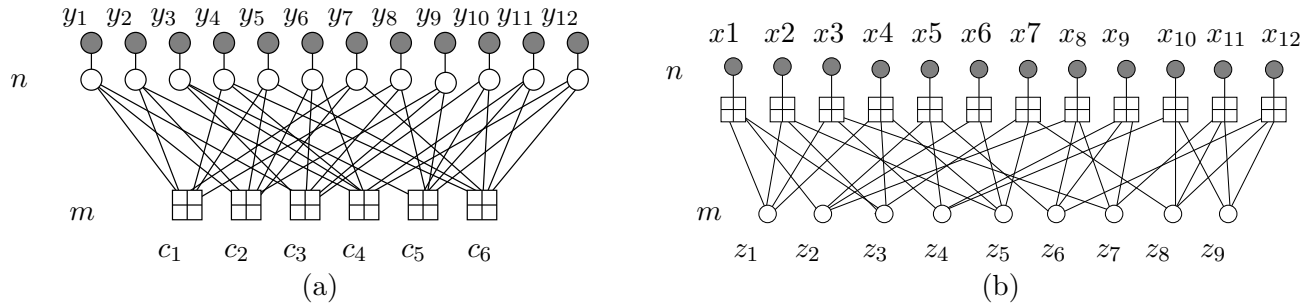


Figure 1. (a) Factor graph representation of a rate R = 0.5 low-density parity check (LDPC) code with bit degree $d_v = 3$ and check degree $d_c' = 6$. (b) Factor graph representation of a rate $R = 0.75$ low-density generator matrix (LDGM) code with check degree $d_c = 3$ and bit degree $d_v = 4$.

Given a binary linear code $\mathbb{C}$, specified by parity check matrix H, the code structure can be captured by a bipartite graph, in which circular nodes (○) represent the binary values x_i (or columns of H), and square nodes (■) represent the parity checks (or rows of H). For instance, Fig. 1(a) shows the factor graph for a rate R = $\frac{1}{2}$ code in parity check form, with $m = 6$ checks acting on $n = 12$ bits. The edges in this graph correspond to 1's in the parity check matrix, and reveal the subset of bits on which each parity check acts. The parity check code in Fig. 1(a) is a regular code with bit degree 3 and check degree 6. Such *low-density* constructions, meaning that both the bit degrees and check degrees remain bounded independently of the block length n, are of most practical use, since they can be efficiently represented and stored, and yield excellent performance under message-passing decoding. In the context of a channel coding problem, the shaded circular nodes at the top of the *low-density parity check* (LDPC) code in panel (a) represent the observed variables y_i received from the noisy channel.

Figure 1(b) shows a binary linear code represented in factor graph form by its generator matrix G. In this dual representation, each codeword $x \in \{0,1\}^n$ is generated by taking the matrix-vector product of the form Gz, where $z \in \{0,1\}^m$ is a sequence of information bits, and $G \in \{0,1\}^{n \times m}$ is the generator matrix. For the code shown in panel (b), the blocklength is $n = 12$, and information sequences are of length $m = 9$, for an overall rate of $R = m/n = 0.75$ in this case. The degrees of the check and variable nodes in the factor graph are $d_c = 3$ and $d_v = 4$ respectively, so that the associated generator matrix G has $d_c = 3$ ones in each row, and $d_v = 4$ ones in each column. When the generator matrix is sparse in this setting, then the resulting code is known as a *low-density generator matrix* (LDGM) code.

2.3 Weight enumerating functions

For future reference, it is useful to define the weight enumerating function of a code. Given a binary linear code of blocklength m, its codewords x have renormalized Hamming weights $w := \frac{\|x\|_1}{m}$ that range in the interval $[0, 1]$. Accordingly, it is convenient to define a function $\mathbb{W}_m : [0, 1] \to \mathbb{R}_+$ that, for each $w \in [0, 1]$, counts the number of codewords of weight w:

$$\mathbb{W}_m(w) \quad := \quad \left| \left\{ x \in \mathbb{C} \;\middle|\; w = \left\lceil \frac{\|x\|_1}{m} \right\rceil \right\} \right|, \tag{2}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. Although it is typically difficult to compute the weight enumerator itself, it is frequently possible to compute (or bound) the *average weight enumerator,* where the expectation is taken over some random ensemble of codes. In particular, our analysis in the sequel makes use of the average weight enumerator of a (d_v, d_c')-regular LDPC code (see Fig. 1(a)), defined as

$$\mathbb{A}_m(w; d_v, d_c') \quad := \quad \frac{1}{m} \log \mathbb{E} \left[\mathbb{W}_m(w) \right], \tag{3}$$

where the expectation is taken over the ensemble of all regular (d_v, d_c')-LDPC codes. For such regular LDPC codes, this average weight enumerator has been extensively studied in previous work [17, 22].

3 Optimality of bounded degree compound constructions

In this section, we describe the compound LDGM/LDPC construction that is the focus of this paper, and describe our main results on their source and channel coding optimality.

3.1 Compound construction

Our main focus is the construction illustrated in Fig. 2, obtained by compounding an LDGM code (top two layers) with an LDPC code (bottom two layers). The code is defined by a factor graph with three layers: at the top, a vector $x \in \{0, 1\}^n$ of codeword bits is connected to a set of n parity checks, which are in turn connected by a sparse generator matrix G to a vector $y \in \{0, 1\}^m$ of information bits in the middle layer. The information bits y are also codewords in an LDPC code, defined by the parity check matrix H connecting the middle and bottom layers.

In more detail, considering first the LDGM component of the compound code, each codeword $x \in \{0, 1\}^n$ in the top layer is connected via the generator matrix $G \in \{0, 1\}^{n \times m}$ to an information sequence $y \in \{0, 1\}^m$ in the middle layer; more specifically, we have the algebraic relation $x = Gy$. Note that this LDGM code has rate $R_G \leq \frac{m}{n}$. Second, turning to the LDPC component of the compound construction, its codewords correspond to a subset of information sequences $y \in \{0, 1\}^m$

in the middle layer. In particular, any valid codeword y satisfies the parity check relation $Hy = 0$, where $H \in \{0,1\}^{m \times k}$ joins the middle and bottom layers of the construction. Overall, this defines an LDPC code with rate $R_H = 1 - \frac{k}{m}$, assuming that H has full row rank.

The overall code $\mathbb{C}$ obtained by concatenating the LDGM and LDPC codes has blocklength n, and rate R upper bounded by $R_G R_H$. In algebraic terms, the code $\mathbb{C}$ is defined as

$$\mathbb{C} \quad := \quad \{x \in \{0,1\}^n \;\; | \;\; x = Gy \quad \text{for some } y \in \{0,1\}^m \text{ such that } \quad Hy = 0\}\,, \tag{4}$$

where all operations are in modulo two arithmetic.

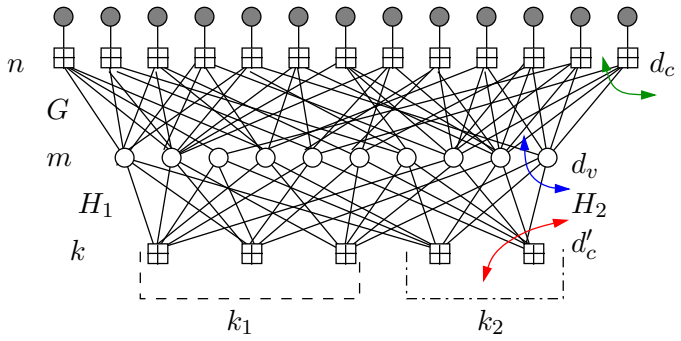


Figure 2. The compound LDGM/LDPC construction analyzed in this paper, consisting of a (n, m) LDGM code over the middle and top layers, compounded with a (m, k) LDPC code over the middle and bottom layers. Codewords $x \in \{0,1\}^n$ are placed on the top row of the construction, and are associated with information bit sequences $z \in \{0,1\}^m$ in the middle layer. The LDGM code over the top and middle layers is defined by a sparse generator matrix $G \in \{0,1\}^{n \times m}$ with at most d_c ones per row. The bottom LDPC over the middle and bottom layers is represented by a sparse parity check matrix $H \in \{0,1\}^{k \times m}$ with d_v ones per column, and d'_c ones per row.

Our analysis in this paper will be performed over random ensembles of compound LDGM/LDPC ensembles. In particular, for each degree triplet (d_c, d_v, d'_c), we focus on the following random ensemble:

(a) For each fixed integer $d_c \geq 4$, the random generator matrix $G \in \{0,1\}^{n \times m}$ is specified as follows: for each of the n rows, we choose d_c positions with replacement, and put a 1 in each of these positions. This procedure yields a random matrix with at most d_c ones per row, since it is possible (although of asymptotically negligible probability for any fixed d_c) that the same position is chosen more than once.

(b) For each fixed degree pair (d_v, d'_c), the random LDPC matrix $H \in \{0,1\}^{k \times m}$ is chosen uniformly at random from the space of all matrices with exactly d_v ones per column, and exactly d'_c ones per row. This ensemble is a standard (d_v, d'_c)-regular LDPC ensemble.

We note that our reason for choosing the check-regular LDGM ensemble specified in step (a) is not that it need define a particularly good code, but rather that it is convenient for theoretical purposes. Interestingly, our analysis shows that the bounded degree d_c check-regular LDGM ensemble, even though it is sub-optimal for both source and channel coding in isolation [28, 29], defines optimal source and channel codes when combined with a bottom LDPC code.

3.2 Main results

Our first main result is on the achievability of the Shannon rate-distortion bound using codes from LDGM/LDPC compound construction with *finite degrees* (d_c, d_v, d'_c). In particular, we make the following claim:

Theorem 1. *Given any pair (R, D) satisfying the Shannon bound, there is a set of finite degrees (d_c, d_v, d'_c) and a code from the associated LDGM/LDPC ensemble with rate R that is a D-good source code (see Definition 1).*

In other work [28, 27], we showed that standard LDGM codes from the check-regular ensemble cannot achieve the rate-distortion bound with finite degrees. As will be highlighted by the proof of Theorem 1 in Section 5, the inclusion of the LDPC lower code in the compound construction plays a vital role in the achievability of the Shannon rate-distortion curve.

Our second main result of this result is complementary in nature to Theorem 1, regarding the achievability of the Shannon channel capacity using codes from LDGM/LDPC compound construction with *finite degrees* (d_c, d_v, d'_c). In particular, we have:

Theorem 2. *For all rate-noise pairs (R, p) satisfying the Shannon channel coding bound $R < 1 - h(p)$, there is a set of finite degrees (d_c, d_v, d'_c) and a code from the associated LDGM/LDPC ensemble with rate R that is a p-good channel code (see Definition 1).*

To put this result into perspective, recall that the overall rate of this compound construction is given by $R = R_G R_H$. Note that an LDGM code on its own (i.e., without the lower LDPC code) is a special case of this construction with $R_H = 1$. However, a standard LDGM of this variety is *not* a good channel code, due to the large number of low-weight codewords. Essentially, the proof of Theorem 2 (see Section 6) shows that using a non-trivial LDPC lower code (with $R_H < 1$) can eliminate these troublesome low-weight codewords.

4 Consequences for coding with side information

We now turn to consideration of the consequences of our two main results for problems of coding with side information. It is well-known from previous work [47] that achieving the information-theoretic limits for these problems requires nested constructions, in which a collection of good source

codes are nested inside a good channel code (or vice versa). Accordingly, we begin in Section 4.1 by describing how our compound construction naturally generates such nested ensembles. In Sections 4.2 and 4.3 respectively, we discuss how the compound construction can be used to achieve the information-theoretic optimum for binary source coding with side information (a version of the Wyner-Ziv problem [45]), and binary information embedding (a version of "dirty paper coding", or the Gelfand-Pinsker problem [19]).

4.1 Nested code structure

The structure of the compound LDGM/LDPC construction lends itself naturally to nested code constructions. In particular, we first partition the set of k lower parity checks into two disjoint subsets K_1 and K_2, of sizes k_1 and k_2 respectively, as illustrated in Fig. 2. Let H_1 and H_2 denote the corresponding partitions of the full parity check matrix $H \in \{0,1\}^{k \times m}$. Now let us set all parity bits in the subset K_2 equal to zero, and consider the LDGM/LDPC code $\mathbb{C}(G, H_1)$ defined by the generator matrix G and the parity check (sub)matrix H_1, as follows

$$\mathbb{C}(G, H_1) \quad := \quad \{x \in \{0,1\}^n \;\; | \;\; x = Gy \quad \text{for some } y \in \{0,1\}^m \text{ such that } \quad H_1 \, y = 0\} \, . \qquad (5)$$

Note that the rate of $\mathbb{C}(G, H_1)$ is given by $\mathrm{R}' = R_G \, R_{H_1}$, which can be suitably adjusted by modifying the LDGM and LDPC rates respectively. Moreover, by applying Theorems 1 and 2, there exist finite choices of degree such that $\mathbb{C}(G, H_1)$ will be optimal for both source and channel coding.

Considering now the remaining k_2 parity bits in the subset K_2, suppose that we set them equal to a fixed binary sequence $\mathbf{m} \in \{0,1\}^{k_2}$. Now consider the code

$$\mathbb{C}(\mathbf{m}) \quad := \quad \left\{ x \in \{0,1\}^n \;\; | \;\; x = Gy \quad \text{for some } y \in \{0,1\}^m \text{ such that } \quad \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \, y = \begin{bmatrix} 0 \\ \mathbf{m} \end{bmatrix} \right\} . \qquad (6)$$

Note that for each binary sequence $\mathbf{m} \in \{0,1\}^{k_2}$, the code $\mathbb{C}(\mathbf{m})$ is a subcode of $\mathbb{C}(G, H_1)$; moreover, the collection of these subcodes forms a disjoint partition as follows

$$\mathbb{C}(G, H_1) \quad = \quad \bigcup_{\mathbf{m} \in \{0,1\}^{k_2}} \mathbb{C}(\mathbf{m}). \qquad (7)$$

Again, Theorems 1 and 2 guarantee that (with suitable degree choices), each of the subcodes $\mathbb{C}(\mathbf{m})$ is optimal for both source and channel coding. Thus, the LDGM/LDPC construction has a natural nested property, in which a good source/channel code—namely $\mathbb{C}(G, H_1)$—is partitioned into a disjoint collection $\{\mathbb{C}(\mathbf{m}), \; \mathbf{m} \in \{0,1\}^{k_1}\}$ of good source/channel codes. We now illustrate how this nested structure can be exploited for coding with side information.

4.2 Source coding with side information

We begin by showing that the compound construction can be used to perform source coding with side information (SCSI).

4.2.1 Problem formulation

Suppose that we wish to compress a symmetric Bernoulli source $S \sim \text{Ber}(\frac{1}{2})$ so as to be able to reconstruct it with Hamming distortion D. As discussed earlier in Section 2, the minimum achievable rate is given by $R(D) = 1 - h(D)$. In the Wyner-Ziv extension of standard lossy compression [45], there is an additional source of side information about S—say in the form $Z = S \oplus W$ where $W \sim \text{Ber}(\delta)$ is observation noise—that is available only at the decoder. See Fig. 3 for a block diagram representation of this problem.

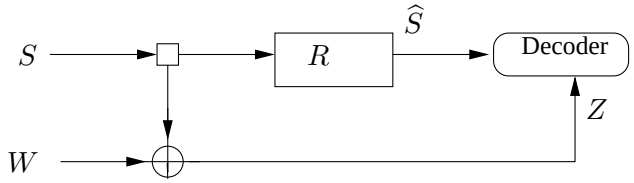


Figure 3. Block diagram representation of source coding with side information (SCSI). A source S is compressed to rate R. The decoder is given the compressed version, and side information $Z = S \oplus W$, and wishes to use $(\widehat{S}, Z)$ to reconstruct the source S up to distortion D.

For this binary version of source coding with side information (SCSI), it is known [2] that the minimum achievable rate takes the form

$$R_{\text{WZ}}(D, p) = \text{l. c. e.} \left\{ h(D * p) - h(D), (p, 0) \right\}, \tag{8}$$

where l. c. e. denotes the lower convex envelope. Note that in the special case $p = \frac{1}{2}$, the side information is useless, so that the Wyner-Ziv rate reduces to classical rate-distortion. In the discussion to follow, we focus only on achieving rates of the form $h(D * p) - h(D)$, as any remaining rates on the Wyner-Ziv curve (8) can be achieved by time-sharing with the point $(p, 0)$.

4.2.2 Coding procedure for SCSI

In order to achieve rates of the form $R = h(D * p) - h(D)$, we use the compound LDGM/LDPC construction, as illustrated in Fig. 2, according to the following procedure.

Step #1, Source coding: The first step is a source coding operation, in which we transform the source sequence S to a quantized representation S. In order to do so, we use the code $\mathbb{C}(G, H_1)$, as defined in equation (5) and illustrated in Fig. 4(a), composed of the generator matrix G and the

parity check matrix H_1. Note that $\mathbb{C}(G, H_1)$, when viewed as a code with blocklength n, has rate $\mathrm{R}_1 := \frac{m\left(1-\frac{k_1}{m}\right)}{n} = \frac{m-k_1}{n}$. Suppose that we choose[1] the middle and lower layer sizes m and k_1 respectively such that

$$\mathrm{R}_1 \;=\; \frac{m-k_1}{n} \;=\; 1 - h\,(D) + \epsilon/2, \tag{9}$$

where $\epsilon > 0$ is arbitrary. For any such choice, Theorem 1 guarantees the existence of finite degrees (d_c, d_v, d_c') such that that $\mathbb{C}(G, H_1)$ is a good D-distortion source code. Consequently, for the specified rate R_1, we can use $\mathbb{C}(G, H_1)$ in order to transform the source to some quantized representation $\widehat{S}$ such that the error $\widehat{S} \oplus S$ has average Hamming weighted bounded by D. Moreover, since $\widehat{S}$ is a codeword of $\mathbb{C}(G, H_1)$, there is some sequence of information bits $\widehat{Y} \in \{0,1\}^m$ such that $\widehat{S} = G\widehat{Y}$ and $H_1\widehat{Y} = 0$.

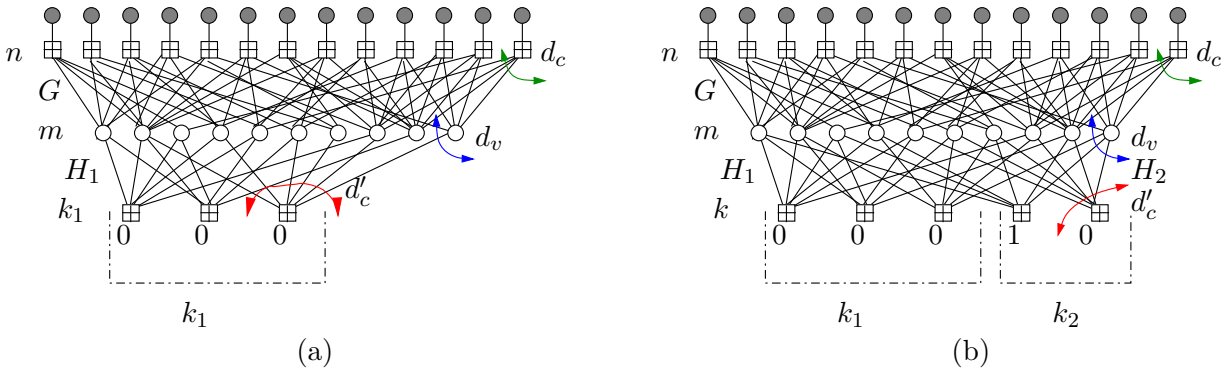


Figure 4. (a) Source coding stage for Wyner-Ziv procedure: the $\mathbb{C}(G, H_1)$, specified by the generator matrix $G \in \{0,1\}^{n\times m}$ and parity check matrix $H_1 \in \{0,1\}^{k_1\times m}$, is used to quantize the source vector $S \in \{0,1\}^n$, thereby obtaining a quantized version $\widehat{S} \in \{0,1\}^n$ and associated vector of information bits $\widehat{Y} \in \{0,1\}^m$, such that $\widehat{S} = G\,\widehat{Y}$ and $H_1\,\widehat{Y} = 0$.

Step #2. Channel coding: Given the output $(\widehat{Y}, \widehat{S})$ of the source coding step, consider the sequence $H_2\widehat{Y} \in \{0,1\}^{k_2}$ of parity bits associated with the parity check matrix H_2. Transmitting this string of parity bits requires rate $R_{\text{trans}} = \frac{k_2}{n}$. Overall, the decoder receives both these k_2 parity bits, as well as the side information sequence $Z = S \oplus W$. Using these two pieces of information, the goal of the decoder is to recover the quantized sequence $\widehat{S}$.

Viewing this problem as one of channel coding, the effective rate of this channel code is $\mathrm{R}_2 = \frac{m-k_1-k_2}{n}$. Note that the side information can be written in the form

$$Z \;=\; S \oplus W \;=\; \widehat{S} \oplus E \oplus W,$$

[1]Note that the choices of m and k_1 need not be unique.

where $E := S \oplus \widehat{S}$ is the quantization noise, and $W \sim \mathrm{Ber}(p)$ is the channel noise. If the quantization noise E were i.i.d. $\mathrm{Ber}(D)$, then the overall effective noise $E \oplus W$ would be i.i.d. $\mathrm{Ber}(D * p)$. (In reality, the quantization noise is not exactly i.i.d. $\mathrm{Ber}(D)$, but it can be shown [47] that it can be treated as such for theoretical purposes.) Consequently, if we choose k_2 such that

$$\mathrm{R}_2 \;=\; \frac{m - k_1 - k_2}{n} \;=\; 1 - h\left(D * p\right) - \epsilon/2, \tag{10}$$

for an arbitrary $\epsilon > 0$, then Theorem 2 guarantees that the decoder will (w.h.p.) be able to recover a codeword corrupted by $(D * p)$-Bernoulli noise.

Summarizing our findings, we state the following:

Corollary 1. *There exist finite choices of degrees (d_c, d_v, d_c') such that the compound LDGM/LDPC construction achieves the Wyner-Ziv bound.*

Proof. With the source coding rate R_1 chosen according to equation (9), the encoder will return a quantization $\widehat{S}$ with average Hamming distance to the source S of at most D. With the channel coding rate R_2 chosen according to equation (10), the decoder can with high probability recover the quantization $\widehat{S}$. The overall transmission rate of the scheme is

$$\begin{aligned} R_{\text{trans}} &= \frac{k_2}{n} \\ &= \frac{m - k_1}{n} - \frac{m - k_1 - k_2}{n} \\ &= \mathrm{R}_1 - \mathrm{R}_2 \\ &= \left(1 - h\left(D\right) + \epsilon/2\right) - \left(1 - h\left(D * p\right) - \epsilon/2\right) \\ &= h\left(D * p\right) - h\left(D\right) + \epsilon. \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we have established that the scheme can achieve rates arbitrarily close to the Wyner-Ziv bound. □

4.3 Channel coding with side information

We now show how the compound construction can be used to perform channel coding with side information (CCSI).

4.3.1 Problem formulation

In the binary information embedding problem, given a specified input vector $V \in \{0,1\}^n$, the channel output $Z \in \{0,1\}^n$ is assumed to take the form

$$Z \;=\; V \oplus S \oplus W, \tag{11}$$

where S is a host signal (not under control of the user), and $W \sim \mathrm{Ber}(p)$ corresponds to channel noise. The encoder is free to choose the input vector $V \in \{0,1\}^n$, subject to an average channel constraint

$$\frac{1}{n}\mathbb{E}\left[\|V\|_1\right] \;\leq\; w, \tag{12}$$

for some parameter $w \in (0, \frac{1}{2}]$. The goal is to use a channel coding scheme that satisfies this constraint (12) so as to maximize the number of possible messages $\mathbf{m}$ that can be reliably communicated. Moreover, We write $V \equiv V_{\mathbf{m}}$ to indicate that each channel input is implicitly identified with some underlying message $\mathbf{m}$. Given the channel output $Z = V_{\mathbf{m}} \oplus S \oplus W$, the goal of the decoder is to recover the embedded message $\mathbf{m}$. The capacity for this binary information embedding problem [2] is given by

$$R_{\mathrm{IE}}(w,p) \;=\; \mathrm{u.\,c.\,e.}\left\{h\left(w\right) - h\left(p\right), (0,0)\right\}, \tag{13}$$

where u. c. e. denotes the upper convex envelope. As before, we focus on achieving rates of the form $h\left(w\right) - h\left(p\right)$, since any remaining points on the curve (13) can be achieved via time-sharing with the $(0,0)$ point.

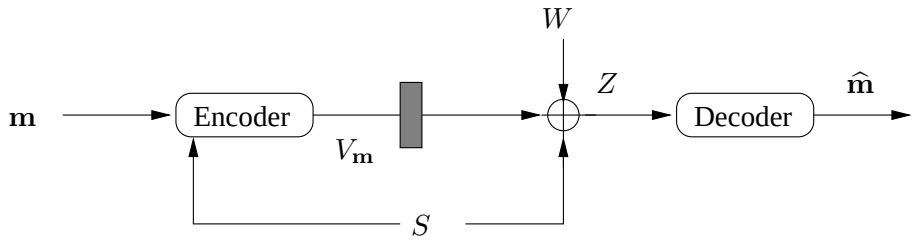


Figure 5. Block diagram representation of channel coding with side information (CCSI). The encoder embeds a message $\mathbf{m}$ into the channel input $V_{\mathbf{m}}$, which is required to satisfy the average channel constraint $\frac{1}{n}\mathbb{E}[\|V_{\mathbf{m}}\|_1] \leq w$. The channel produces the output $Z = V_{\mathbf{m}} \oplus S \oplus W$, where S is a host signal known only to the encoder, and $W \sim \mathrm{Ber}(p)$ is channel noise. Given the channel output Y, the decoder outputs an estimate $\widehat{\mathbf{m}}$ of the embedded message.

4.3.2 Coding procedure for CCSI

In order to achieve rates of the form $R = h(w) - h(p)$, we again use the compound LDGM/LDPC construction in Fig. 2, now according to the following two step procedure.

Step #1: Source coding: The goal of the first stage is to embed the message into the transmitted signal V via a quantization process. In order to do so, we use the code illustrated in Fig. 6(a), specified by the generator matrix G and parity check matrices H_1 and H_2. The set K_1 of k_1 parity bits associated with the check matrix H_1 remain fixed to zero throughout the scheme. On the other hand, we use the remaining k_2 lower parity bits associated with H_2 to specify a particular message $\mathbf{m} \in \{0,1\}^{k_2}$ that the decoder would like to recover. In algebraic terms, the resulting code $\mathbb{C}(\mathbf{m})$ has the form

$$\mathbb{C}(\mathbf{m}) \;:=\; \left\{ x \in \{0,1\}^n \;\mid\; x = Gy \quad \text{for some } y \in \{0,1\}^m \text{ such that } \quad \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} y = \begin{bmatrix} 0 \\ \mathbf{m} \end{bmatrix} \right\}. \tag{14}$$

Since the encoder has access to host signal S, it may use this code $\mathbb{C}(\mathbf{m})$ in order to quantize the host signal. After doing so, the encoder has a quantized signal $\widehat{S}_{\mathbf{m}} \in \{0,1\}^n$ and an associated sequence $\widehat{Y}_{\mathbf{m}} \in \{0,1\}^m$ of information bits such that $\widehat{S}_{\mathbf{m}} = G\widehat{Y}_{\mathbf{m}}$. Note that the quantized signal $(\widehat{Y}_{\mathbf{m}}, \widehat{S}_{\mathbf{m}})$ specifies the message $\mathbf{m}$ in an implicit manner, since $\mathbf{m} = H_2 \widehat{Y}_{\mathbf{m}}$ by construction of the code $\mathbb{C}(\mathbf{m})$.

Now suppose that we choose n, m and k such that

$$\mathrm{R}_1 \;=\; \frac{m - k_1 - k_2}{n} \;=\; 1 - h(w) + \epsilon/2 \tag{15}$$

for some $\epsilon > 0$, then Theorem 1 guarantees that there exist finite degrees (d_c, d_v, d_c') such that the resulting code is a good w-distortion source code. Otherwise stated, we are guaranteed that w.h.p, the quantization error $E := S \oplus \widehat{S}$ has average Hamming weight upper bounded by wn. Consequently, we may set the channel input V equal to the quantization noise ($V = E$), thereby ensuring that the average channel constraint (12) is satisfied.

Step #2, Channel coding: In the second phase, the decoder is given a noisy channel observation of the form

$$Z \;=\; E \oplus S \oplus W \;=\; \widehat{S} \oplus W, \tag{16}$$

and its task is to recover $\widehat{S}$. In terms of the code architecture, the k_1 lower parity bits remain set to zero; the remaining k_2 parity bits, which represent the message $\mathbf{m}$, are unknown to the coder. The resulting code, as illustrated illustrated in Fig. 6(b), can be viewed as channel code with effective

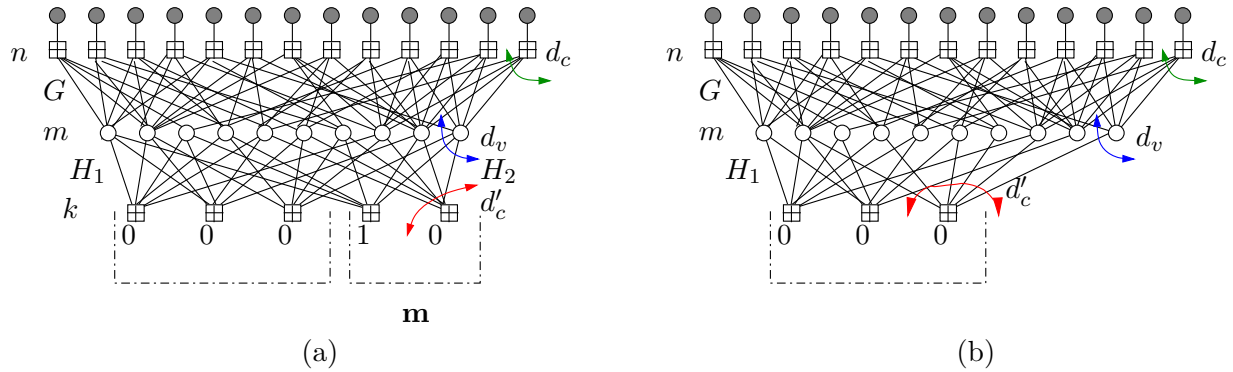


Figure 6. (a) Source coding step for binary information embedding. The message $\mathbf{m} \in \{0,1\}^{k_2}$ specifies a particular coset; using this particular source code, the host signal S is compressed to $\widehat{S}$, and the quantization error $E = S \oplus \widehat{S}$ is transmitted over the constrained channel. (b) Channel coding step for binary information embedding. The decoder receives $Z = \widehat{S} \oplus W$ where $W \sim \mathrm{Ber}(p)$ is channel noise, and seeks to recover $\widehat{S}$, and hence the embedded message $\mathbf{m}$ specifying the coset.

rate $\frac{m-k_1}{n}$. Now suppose that we choose k_1 such that the effective code used by the decoder has rate

$$\mathrm{R}_2 \quad = \quad \frac{m-k_1}{n} = 1 - h\left(p\right) - \epsilon/2, \tag{17}$$

for some $\epsilon > 0$. Since the channel noise W is $\mathrm{Ber}(p)$ and the rate R_2 chosen according to (17), Theorem 2 guarantees that the decoder will w.h.p. be able to recover the pair $\widehat{S}$ and $\widehat{Y}$. Moreover, by design of the quantization procedure, we have the equivalence $\mathbf{m} = H_2 \widehat{Y}$ so that a simple syndrome-forming procedure allows the decoder to recover the hidden message.

Summarizing our findings, we state the following:

Corollary 2. *There exist finite choices of degrees (d_c, d_v, d'_c) such that the compound LDGM/LDPC construction achieves the binary information embedding (Gelfand-Pinsker) bound.*

Proof. With the source coding rate R_1 chosen according to equation (15), the encoder will return a quantization $\widehat{S}$ of the host signal S with average Hamming distortion upper bounded by w. Consequently, transmitting the quantization error $E = S \oplus \widehat{S}$ will satisfy the average channel constraint (12). With the channel coding rate R_2 chosen according to equation (17), the decoder can with high probability recover the quantized signal $\widehat{S}$, and hence the message $\mathbf{m}$. Overall, the scheme allows a total of 2^{k_2} distinct messages to be embedded, so that the effective information

embedding rate is

$$\begin{aligned} R_{\text{trans}} &= \frac{k_2}{n} \\ &= \frac{m-k_1}{n} - \frac{m-k_1-k_2}{n} \\ &= \mathrm{R}_2 - \mathrm{R}_1 \\ &= (1-h(p)-\epsilon/2) - (1-h(w)+\epsilon/2) \\ &= h(w) - h(p) + \epsilon, \end{aligned}$$

for some $\epsilon > 0$. Thus, we have shown that the proposed scheme achieves the binary information embedding bound (13). □

5 Proof of source coding optimality

This section is devoted to the proof of the previously stated Theorem 1 on the source coding optimality of the compound construction.

5.1 Set-up

In establishing a rate-distortion result such as Theorem 1, perhaps the most natural focus is the random variable

$$d_n(S,\mathbb{C}) := \frac{1}{n}\min_{x\in\mathbb{C}} \|x-S\|_1, \tag{18}$$

corresponding to the (renormalized) minimum Hamming distance from a random source sequence $S \in \{0,1\}^n$ to the nearest codeword in the code $\mathbb{C}$. Rather than analyzing this random variable directly, our proof of Theorem 1 proceeds indirectly, by studying an alternative random variable.

Given a binary linear code with N codewords, let $i = 0, 1, 2, \ldots, N-1$ be indices for the different codewords. We say that a codeword X^i is *distortion D-good* for a source sequence S if the Hamming distance $\|X^i \oplus S\|_1$ is at most Dn. We then set the indicator random variable $Z^i(D) = 1$ when codeword X^i is distortion D-good. With these definitions, our proof is based on the following random variable:

$$T_n(S,\mathbb{C};D) := \sum_{i=0}^{N-1} Z^i(D). \tag{19}$$

Note that $T_n(S,\mathbb{C};D)$ simply counts the number of codewords that are distortion D-good for a source sequence S. Moreover, for all distortions D, the random variable $T_n(S,\mathbb{C};D)$ is linked to

$d_n(S, \mathbb{C})$ via the equivalence

$$\mathbb{P}[T_n(S, \mathbb{C}; D) > 0] \quad = \quad \mathbb{P}[d_n(S, \mathbb{C}) \leq D]. \tag{20}$$

Throughout our analysis of $\mathbb{P}[T_n(S, \mathbb{C}; D) > 0]$, we carefully track only its exponential behavior. More precisely, the analysis to follow will establish an inverse polynomial lower bound of the form $\mathbb{P}[T_n(S, \mathbb{C}; D) > 0] \geq 1/f(n)$ where $f(\cdot)$ collects various polynomial factors. The following concentration result establishes that the polynomial factors in these bounds can be ignored:

Lemma 1 (Sharp concentration)**.** *Suppose that for some target distortion D, we have*

$$\mathbb{P}[T_n(S, \mathbb{C}; D) > 0] \quad \geq \quad 1/f(n), \tag{21}$$

where $f(\cdot)$ is a polynomial function satisfying $\log f(n) = o(n)$. Then for all $\epsilon > 0$, there exists a fixed code $\bar{\mathbb{C}}$ *of sufficiently large blocklength n such that $\mathbb{E}[d_n(S; \bar{\mathbb{C}})] \leq D + \epsilon$.*

Proof. Let us denote the random code $\mathbb{C}$ as $(\mathbb{C}_1, \mathbb{C}_2)$, where $\mathbb{C}_1$ denotes the random LDGM top code, and $\mathbb{C}_2$ denotes the random LDPC bottom code. Throughout the analysis, we condition on some fixed LDPC bottom code, say $\mathbb{C}_2 = \bar{\mathbb{C}}_2$. We begin by showing that the random variable $(d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2)$ is sharply concentrated. In order to do so, we construct a vertex-exposure martingale [33] of the following form. Consider a fixed sequential labelling $\{1, \ldots, n\}$ of the top LDGM checks, with check i associated with source bit S_i. We reveal the check and associated source bit in a sequential manner for each $i = 1, \ldots, n$, and so define a sequence of random variables $\{U_0, U_1, \ldots, U_n\}$ via $U_0 := \mathbb{E}[d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2]$, and

$$U_i \quad := \quad \mathbb{E}\left[d_n(S, \mathbb{C}) \;\mid\; S_1, \ldots, S_i, \bar{\mathbb{C}}_2\right], \qquad i = 1, \ldots, n. \tag{22}$$

By construction, we have $U_n = (d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2)$. Moreover, this sequence satisfies the following bounded difference property: adding any source bit S_i and the associated check in moving from U_{i-1} to U_i can lead to a (renormalized) change in the minimum distortion of at most $c_i = 1/n$. Consequently, by applying Azuma's inequality [1], we have, for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|(d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2) - \mathbb{E}[d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2]\right| \geq \epsilon\right] \quad \leq \quad \exp\left(-n\epsilon^2\right). \tag{23}$$

Next we observe that our assumption (21) of inverse polynomial decay implies that, for at least one bottom code $\bar{\mathbb{C}}_2$,

$$\mathbb{P}[d_n(S, \mathbb{C}) \leq D \mid \bar{\mathbb{C}}_2] \;=\; \mathbb{P}[T_n(S, \mathbb{C}; D) > 0 \mid \bar{\mathbb{C}}_2] \quad \geq \quad 1/g(n), \tag{24}$$

for some subexponential function g. Otherwise, there would exist some $\alpha > 0$ such that

$$\mathbb{P}[T_n(S, \mathbb{C}; D) > 0 \mid \bar{\mathbb{C}}_2] \leq \exp(-n\alpha)$$

for all choices of bottom code $\bar{\mathbb{C}}_2$, and taking averages would violate our assumption (21).

Finally, we claim that the concentration result (23) and inverse polynomial bound (24) yield the result. Indeed, if for some $\epsilon > 0$, we had $D < \mathbb{E}[d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2] - \epsilon$, then the concentration bound (23) would imply that the probability

$$\begin{aligned} \mathbb{P}[d_n(S, \mathbb{C}) \leq D \mid \bar{\mathbb{C}}_2] &\leq \mathbb{P}[d_n(S, \mathbb{C}) \leq \mathbb{E}[d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2] - \epsilon \mid \bar{\mathbb{C}}_2] \\ &\leq \mathbb{P}\left[\left|(d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2) - \mathbb{E}[d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2]\right| \geq \epsilon\right] \end{aligned}$$

decays exponentially, which would contradict the inverse polynomial bound (24) for sufficiently large n. Thus, we have shown that assumption (21) implies that for all $\epsilon > 0$, there exists a sufficiently large n and fixed bottom code $\bar{\mathbb{C}}_2$ such that $\mathbb{E}[d_n(S, \mathbb{C}) \mid \bar{\mathbb{C}}_2] \leq D + \epsilon$. If the average over LDGM codes $\mathbb{C}_1$ satisfies this bound, then at least one choice of LDGM top code must also satisfy it, whence we have established that there exists a fixed code $\bar{\mathbb{C}}$ such that $\mathbb{E}[d_n(S; \bar{\mathbb{C}})] \leq D + \epsilon$, as claimed. □

5.2 Moment analysis

In order to analyze the probability $\mathbb{P}[T_n(S, \mathbb{C}; D) > 0]$, we make use of the moment bounds given in the following elementary lemma:

Lemma 2 (Moment methods)**.** *Given any random variable N taking non-negative integer values, there holds*

$$\frac{(\mathbb{E}[N])^2}{\mathbb{E}[N^2]} \overset{(a)}{\leq} \mathbb{P}[N > 0] \overset{(b)}{\leq} \mathbb{E}[N]. \tag{25}$$

Proof. The upper bound (b) is an immediate consequence of Markov's inequality, whereas the lower bound (a) follows by applying the Cauchy-Schwarz inequality [20] as follows

$$(\mathbb{E}[N])^2 = \big(\mathbb{E}\big[N\, \mathbb{I}[N > 0]\big]\big)^2 \leq \mathbb{E}[N^2]\, \mathbb{E}\left[\mathbb{I}^2[N > 0]\right] = \mathbb{E}[N^2]\, \mathbb{P}[N > 0].$$

□

The remainder of the proof consists in applying these moment bounds to the random variable $T_n(S, \mathbb{C}; D)$, in order to bound the probability $\mathbb{P}[T_n(S, \mathbb{C}; D) > 0]$. We begin by computing the first moment:

Lemma 3 (First moment)**.** *For any code with rate R, the expected number of D-good codewords scales exponentially as*

$$\frac{1}{n}\log \mathbb{E}[T_n] \;\; = \;\; [\mathrm{R} - (1 - h\,(D))] \;\pm\; o(1). \tag{26}$$

Proof. First, by linearity of expectation $\mathbb{E}[T_n] = \sum_{i=0}^{2^{n\mathrm{R}}-1} \mathbb{P}[Z^i(D) = 1] \;=\; 2^{n\mathrm{R}}\mathbb{P}[Z^0(D) = 1]$, where we have used symmetry of the code construction to assert that $\mathbb{P}[Z^i(D) = 1] = \mathbb{P}[Z^0(D) = 1]$ for all indices i. Now the event $\{Z^0(D) = 1\}$ is equivalent to an i.i.d Bernoulli($\frac{1}{2}$) sequence of length n having Hamming weight less than or equal to Dn. By standard large deviations theory (either Sanov's theorem [11], or direct asymptotics of binomial coefficients), we have

$$\frac{1}{n}\log \mathbb{P}[Z^0(D) = 1] \;\; = \;\; 1 - h\,(D) \;\pm\; o(1),$$

which establishes the claim.

□

Unfortunately, however, the first moment $\mathbb{E}[T_n]$ need not be representative of typical behavior of the random variable T_n, and hence overall distortion performance of the code. As a simple illustration, consider an imaginary code consisting of $2^{n\mathrm{R}}$ copies of the all-zeroes codeword. Even for this "code", as long as $\mathrm{R} > 1 - h\,(D)$, the expected number of distortion-D optimal codewords grows exponentially. Indeed, although $T_n = 0$ for almost all source sequences, for a small subset of source sequences (of probability mass $\approx 2^{-n\,[1-h(D)]}$), the random variable T_n takes on the enormous value 2^{nR}, so that the first moment grows exponentially. However, the average distortion incurred by using this code will be ≈ 0.5 for any rate, so that the first moment is entirely misleading. In order to assess the representativeness of the first moment, one needs to ensure that it is of essentially the same order as the variance, hence the comparison involved in the second moment bound (25)(a).

5.3 Second moment analysis

Our analysis of the second moment begins with the following alternative representation:

Lemma 4.

$$\mathbb{E}[T_n^2(D)] \;=\; \mathbb{E}[T_n(D)] \left(1 + \Big\{ \sum_{j \neq 0} \mathbb{P}[Z^j(D) = 1 \;\mid\; Z^0(D) = 1] \Big\}\right). \tag{27}$$

Based on this lemma, proved in Appendix C, we see that the key quantity to control is the conditional probability $\mathbb{P}[Z^j(D) = 1 \;\mid\; Z^0(D) = 1]$. It is this *overlap probability* that differentiates the low-density codes of interest here from the unstructured codebooks used in classical random coding

arguments.[2] For a low-density graphical code, the dependence between the events $\{Z^j(D) = 1\}$ and $\{Z^0(D) = 1\}$ requires some analysis.

Before proceeding with this analysis, we require some definitions. Recall our earlier definition (3) of the average weight enumerator associated with an (d_v, d_c') LDPC code, denoted by $\mathbb{A}_m(w)$. Moreover, let us define for each $w \in [0, 1]$ the probability

$$\mathbb{Q}(w; D) \quad := \quad \mathbb{P}\left[\|X(w) \oplus S\|_1 \leq Dn \;\mid\; \|S\|_1 \leq Dn\right], \tag{28}$$

where the quantity $X(w) \in \{0, 1\}^n$ denotes a randomly chosen codeword, conditioned on its underlying length-m information sequence having Hamming weight $\lceil wm \rceil$. As shown in Lemma 9 (see Appendix A), the random codeword $X(w)$ has i.i.d. Bernoulli elements with parameter

$$\delta^*(w; d_c) \quad = \quad \frac{1}{2}\left[1 - (1 - 2\,w)^{d_c}\right]. \tag{29}$$

With these definitions, we now break the sum on the RHS of equation (27) into m terms, indexed by $t = 1, 2, \ldots, m$, where term t represents the contribution of a given non-zero information sequence $y \in \{0, 1\}^m$ with (Hamming) weight t. Doing so yields

$$\begin{aligned}
\sum_{j \neq 0} \mathbb{P}[Z^j(D) = 1 \mid Z^0(D) = 1] &= \sum_{t=1}^{m} \mathbb{A}_m(t/m)\, \mathbb{Q}(t/m; D) \\
&\leq m \max_{1 \leq t \leq m} \{\mathbb{A}_m(t/m)\; \mathbb{Q}(t/m; D)\} \\
&\leq m \max_{w \in [0,1]} \{\mathbb{A}_m(w)\; \mathbb{Q}(w; D)\}.
\end{aligned}$$

Consequently, we need to control both the LDPC weight enumerator $\mathbb{A}_m(w)$ and the probability $\mathbb{Q}(w; D)$ over the range of possible fractional weights $w \in [0, 1]$.

5.4 Bounding the overlap probability

The following lemma, proved in Appendix D, provides a large deviations bound on the probability $\mathbb{Q}(w; D)$.

Lemma 5. *For each $w \in [0, 1]$, we have*

$$\frac{1}{n} \log \mathbb{Q}(w; D) \quad \leq \quad F(\delta^*(w; d_c); D) + o(1), \tag{30}$$

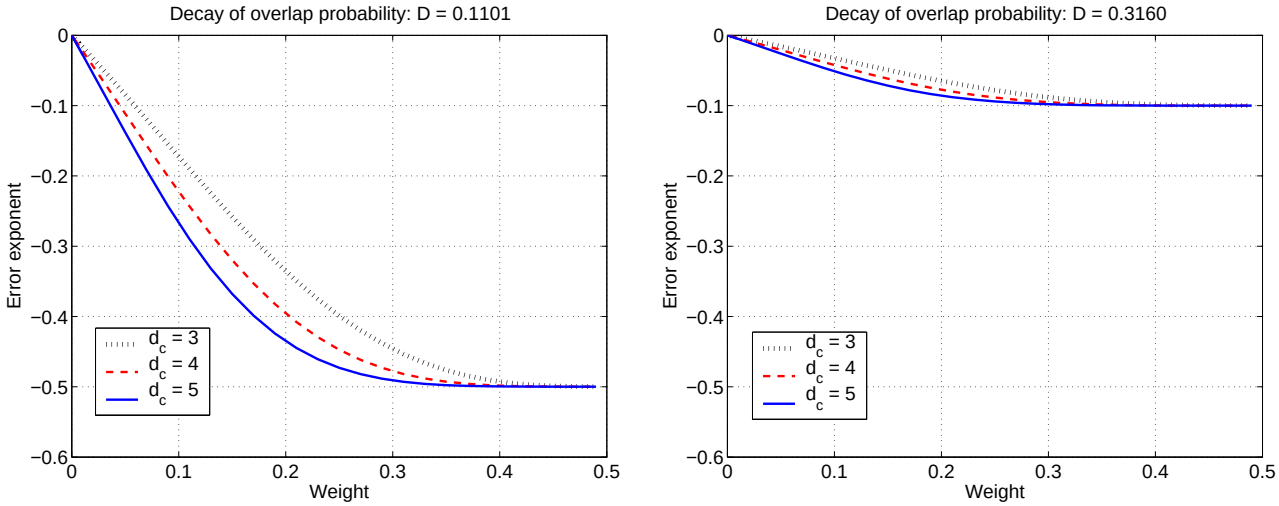
[2] In the latter case, codewords are chosen independently from some ensemble, so that the overlap probability is simply equal to $\mathbb{P}[Z^j(D) = 1]$. Thus, for the simple case of unstructured random coding, the second moment bound actually provides the converse to Shannon's rate-distortion theorem for the symmetric Bernoulli source.

where for each $t \in (0, \frac{1}{2}]$ *and* $D \in (0, \frac{1}{2}]$*, the error exponent is given by*

$$F(t; D) \quad := \quad D \log \left[(1-t)e^{\lambda^*} + t \right] + (1-D) \log \left[(1-t) + te^{\lambda^*} \right] - \lambda^* D. \tag{31}$$

Here $\lambda^* := \log \left[\frac{-b + \sqrt{b^2 - 4ac}}{2a} \right]$*, where* $a := t\,(1-t)\,(1-D)$*,* $b := (1-2D)t^2$*, and* $c := -t\,(1-t)\,D$*.*

In general, for any $D \in (0, \frac{1}{2}]$, the function $F(\,\cdot\,; D)$ has the following properties. At $t = 0$, it achieves its maximum $F(0\,; D) = 0$, and then is strictly decreasing on the interval $(0, \frac{1}{2}]$, approaching its minimum value $-[1 - h\,(D)]$ as $t \to \frac{1}{2}$. Figure 7 illustrates the form of the function $F(\delta^*(\omega; d_c); D)$ for two different values of distortion D, and for degrees $d_c \in \{3, 4, 5\}$. Note that



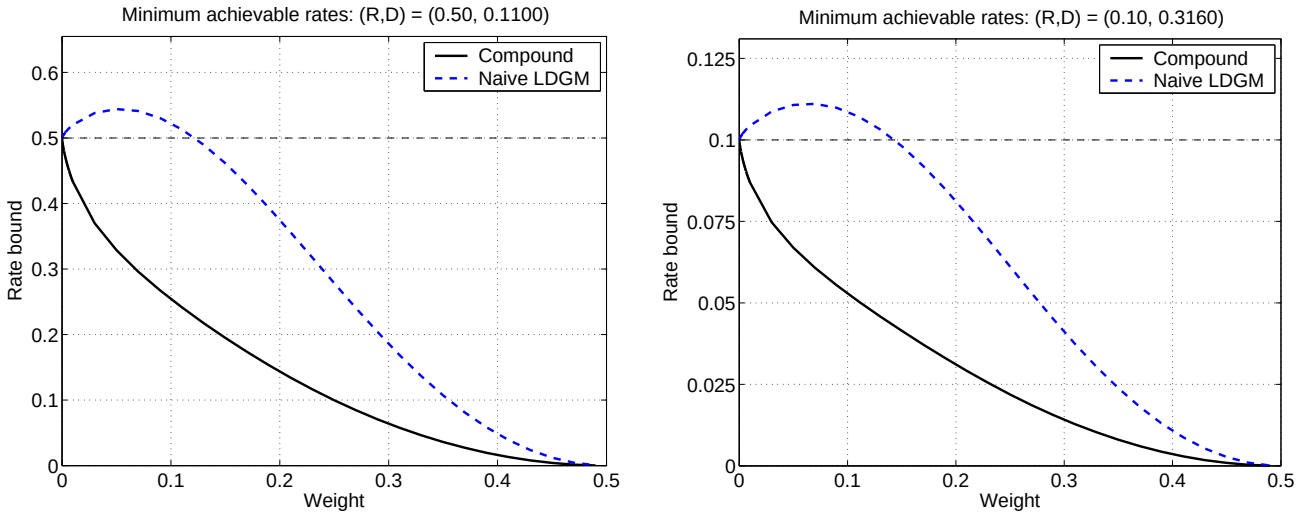
(a) (b)

Figure 7. Plot of the upper bound (30) on the overlap probability $\frac{1}{n} \log \mathbb{Q}(w; D)$ for different choices of the degree d_c, and distortion probabilities. (a) Distortion $D = 0.1100$. (b) Distortion $D = 0.3160$.

increasing d_c causes $F(\delta^*(\omega; d_c); D)$ to approach its minimum $-[1 - h\,(D)]$ more rapidly.

We are now equipped to establish the form of the effective rate-distortion function for any compound LDGM/LDPC ensemble. Substituting the alternative form of $\mathbb{E}[T_n^2]$ from equation (27) into the second moment lower bound (25) yields

$$\begin{aligned}
\frac{1}{n} \log \mathbb{P}[T_n(D) > 0] \quad &\geq \quad \frac{1}{n} \left[\log \mathbb{E}[T_n(D)] - \log \left\{ 1 + \sum_{j \neq 0} \mathbb{P}[Z^j(D) = 1 \;\mid\; Z^0(D) = 1] \right\} \right] \\
&\geq \quad R - (1 - h\,(D)) - \max_{w \in [0,1]} \left\{ \frac{1}{n} \log \mathbb{A}_m(w) + \frac{1}{n} \log \mathbb{Q}(w; D) \right\} - o(1) \\
&\geq \quad R - (1 - h\,(D)) - \max_{w \in [0,1]} \left\{ R\, \frac{1}{R_H} \frac{\log \mathbb{A}_m(w)}{m} + F(\,\delta^*(w; d_c), D) \right\} - o(1), \qquad (32)
\end{aligned}$$



(a) (b)

Figure 8. Plot of the function defining the lower bound (33) on the minimum achievable rate for a specified distortion. Shown are curves with LDGM top degree $d_c = 4$, comparing the uncoded case (no bottom code, dotted curve) to a bottom $(4, 6)$ LDPC code (solid line). (a) Distortion $D = 0.1100$. (b) Distortion $D = 0.3160$.

where the last step follows by applying the upper bound on $\mathbb{Q}$ from Lemma 5, and the relation $m = R_G n = \frac{R}{R_H} n$. Now letting $B(w; d_v, d_c')$ be any upper bound on the log of average weight enumerator $\frac{\log \mathbb{A}_m(w)}{m}$, we can then conclude that $\frac{1}{n} \log \mathbb{P}[T_n(D) > 0]$ is asymptotically non-negative for all rate-distortion pairs (R, D) satisfying

$$R \geq \max_{w \in [0,1]} \left[\frac{1 - h(D) + F(\delta^*(w; d_c), D)}{1 - \frac{B(w; d_v, d_c')}{R_H}} \right]. \tag{33}$$

Figure 8 illustrates the behavior of the RHS of equation (33), whose maximum defines the effective rate-distortion function, for the case of LDGM top degree $d_c = 4$. Panels (a) and (b) show the cases of distortion $D = 0.1100$ and $D = 0.3160$ respectively, for which the respective Shannon rates are $R = 0.50$ and $R = 0.10$. Each panel shows two plots, one corresponding the case of uncoded information bits (a naive LDGM code), and the other to using a rate $R_H = 2/3$ LDPC code with degrees $(d_v, d_c) = (4, 6)$. In all cases, the minimum achievable rate for the given distortion is obtained by taking the maximum for $w \in [0, 0.5]$ of the plotted function. For any choices of D, the plotted curve is equal to the Shannon bound $R_{\text{Sha}} = 1 - h(D)$ at $w = 0$, and decreases to 0 for $w = \frac{1}{2}$.

Note the dramatic difference between the uncoded and compound constructions (LDPC-coded). In particular, for both settings of the distortion ($D = 0.1100$ and $D = 0.3160$), the uncoded curves rise from their initial values to maxima *above* the Shannon limit (dotted horizontal line). Con-

sequently, the minimum required rate using these constructions lies strictly above the Shannon optimum. The compound construction curves, in contrast, decrease monotonically from their maximum value, achieved at $w = 0$ and corresponding to the Shannon optimum. In the following section, we provide an analytical proof of the fact that for any distortion $D \in [0, \frac{1}{2})$, it is always possible to choose finite degrees such that the compound construction achieves the Shannon optimum.

5.5 Finite degrees are sufficient

In order to complete the proof of Theorem 1, we need to show that for all rate-distortion pairs (R, D) satisfying the Shannon bound, there exist LDPC codes with finite degrees (d_v, d_c') and a suitably large but finite top degree d_c such that the compound LDGM/LDPC construction achieves the specified (R, D).

Our proof proceeds as follows. Recall that in moving from equation (32) to equation (33), we assumed a bound on the average weight enumerator $\mathbb{A}_m$ of the form

$$\frac{1}{m} \log \mathbb{A}_m(w) \quad \leq \quad B(w; d_v, d_c') + o(1). \tag{34}$$

For compactness in notation, we frequently write $B(w)$, where the dependence on the degree pair (d_v, d_c') is understood implicitly. In the following paragraph, we specify a set of conditions on this bounding function B, and we then show that under these conditions, there exists a finite degree d_c such that the compound construction achieves specified rate-distortion point. In Appendix F, we then prove that the weight enumerator of standard regular LDPC codes satisfies the assumptions required by our analysis.

Assumptions on weight enumerator bound We require that our bound B on the weight enumerator satisfy the following conditions:

A1: the function B is symmetric around $\frac{1}{2}$, meaning that $B(w) = B(1 - w)$ for all $w \in [0, 1]$.

A2: the function B is twice differentiable on $(0, 1)$ with $B'(\frac{1}{2}) = 0$ and $B''(\frac{1}{2}) < 0$.

A3: the function B achieves its unique optimum at $w = \frac{1}{2}$, where $B(\frac{1}{2}) = R_H$.

A4: there exists some $\epsilon_1 > 0$ such that $B(w) < 0$ for all $w \in (0, \epsilon_1)$, meaning that the ensemble has linear minimum distance.

In order to establish our claim, it suffices to show that for all (R, D) such that $R > 1 - h(D)$, there exists a finite choice of d_c such that

$$\max_{w \in [0,1]} \underbrace{\left\{ R \, \frac{B(w)}{R_H} + F(\delta^*(w; d_c), D) \right\}}_{K(w; d_c)} \quad \leq \quad R - [1 - h(D)] \; := \; \Delta \tag{35}$$

Restricting to even d_c ensures that the function F is symmetric about $w = \frac{1}{2}$; combined with assumption A2, this ensures that K is symmetric around $\frac{1}{2}$, so that we may restrict the maximization to $[0, \frac{1}{2}]$ without loss of generality. Our proof consists of the following steps:

(a) We first prove that there exists an $\epsilon_1 > 0$, independent of the choice of d_c, such that $K(w; d_c) \leq \Delta$ for all $w \in [0, \epsilon_1]$.

(b) We then prove that there exists $\epsilon_2 > 0$, again independent of the choice of d_c, such that $K(w; d_c) \leq \Delta$ for all $w \in [\frac{1}{2} - \epsilon_2, \frac{1}{2}]$.

(c) Finally, we specify a sufficiently large but finite degree d_c^* that ensures the condition $K(w; d_c^*) \leq \Delta$ for all $w \in [\epsilon_1, \epsilon_2]$.

5.5.1 Step A

By assumption A4 (linear minimum distance), there exists some $\epsilon_1 > 0$ such that $B(w) \leq 0$ for all $w \in [0, \epsilon_1]$. Since $F(\delta^*(w; d_c); D) \leq 0$ for all w, we have $K(w; d_c) \leq 0 < \Delta$ in this region. Note that ϵ_1 is independent of d_c, since it specified entirely by the properties of the bottom code.

5.5.2 Step B

For this step of the proof, we require the following lemma on the properties of the function F:

Lemma 6. *For all choices of even degrees $d_c \geq 4$, the function $G(w; d_c) = F(\delta^*(w; d_c), D)$ is differentiable in a neighborhood of $w = \frac{1}{2}$, with*

$$G(\frac{1}{2}; d_c) = -[1 - h(D)], \qquad G'(\frac{1}{2}; d_c) = 0, \qquad \textit{and} \qquad G''(\frac{1}{2}; d_c) = 0. \tag{36}$$

See Appendix E for a proof of this claim. Next observe that we have the uniform bound $G(w; d_c) \leq G(w; 4)$ for all $d_c \geq 4$ and $w \in [0, \frac{1}{2}]$. This follows from the fact that $F(u; D)$ is decreasing in u, and that $\delta^*(w; 4) \leq \delta^*(w; d_c)$ for all $d_c \geq 4$ and $w \in [0, \frac{1}{2}]$. Since B is independent of d_c, this implies that $K(w; d_c) \leq K(w; 4)$ for all $w \in [0, \frac{1}{2}]$. Hence it suffices to set $d_c = 4$, and show that $K(w; 4) \leq \Delta$ for all $w \in [\frac{1}{2} - \epsilon_2, \frac{1}{2}]$. Using Lemma 6, Assumption A2 concerning the

derivatives of B, and Assumption A4 (that $B(\frac{1}{2}) = R_H$), we have

$$\begin{aligned} K(\frac{1}{2};4) &= R - [1 - h\,(D)] \;=\; \Delta, \\ K'(\frac{1}{2};4) &= \frac{R\;B'(\frac{1}{2})}{R_H} + G'(\frac{1}{2};4) \;=\; 0, \qquad \text{and} \\ K''(\frac{1}{2};4) &= \frac{R\;B''(\frac{1}{2})}{R_H} + G''(\frac{1}{2};4) \;=\; \frac{R\;B''(\frac{1}{2})}{R_H} < 0. \end{aligned}$$

By the continuity of K'', the second derivative remains negative in a region around $\frac{1}{2}$, say for all $w \in [\frac{1}{2} - \epsilon_2, \frac{1}{2}]$ for some $\epsilon_2 > 0$. Then, for all $w \in [\frac{1}{2} - \epsilon_2, \frac{1}{2}]$, we have for some $\widetilde{w} \in [w, \frac{1}{2}]$ the second order expansion

$$\begin{aligned} K(w;4) &= K(\frac{1}{2};4) + K'(\frac{1}{2};4)(w - \frac{1}{2}) + \frac{1}{2}K(\widetilde{w};4)\;\left(w - \frac{1}{2}\right)^2 \\ &= \Delta + \frac{1}{2}K(\widetilde{w};4)\;\left(w - \frac{1}{2}\right)^2 \;\leq\; \Delta. \end{aligned}$$

Thus, we have established that there exists an $\epsilon_2 > 0$, independent of the choice of d_c, such that for all even $d_c \geq 4$, we have

$$K(w;d_c) \;\leq\; K(w,4) \;\leq\; \Delta \qquad \text{for all } w \in [\tfrac{1}{2} - \epsilon_2, \tfrac{1}{2}]. \tag{37}$$

5.5.3 Step C

Finally, we need to show that $K(w;d_c) \leq \Delta$ for all $w \in [\epsilon_1, \epsilon_2]$. From assumption A3 and the continuity of B, there exists some $\rho(\epsilon_2) > 0$ such that

$$B(w) \;\leq\; R_H\;[1 - \rho(\epsilon_2)] \qquad \text{for all } w \leq \tfrac{1}{2} - \epsilon_2. \tag{38}$$

From Lemma 6, $\lim_{u \to \frac{1}{2}} F(u;D) = F(\frac{1}{2};D) \;=\; -\,[1 - h\,(D)]$. Moreover, as $d_c \to +\infty$, we have $\delta^*(\epsilon_1;d_c) \to \frac{1}{2}$. Therefore, for any $\epsilon_3 > 0$, there exists a finite degree d_c^* such that

$$F(\delta^*(\epsilon_1;d_c^*);D) \;\leq\; -\,[1 - h\,(D)] + \epsilon_3.$$

Since F is non-increasing in w, we have $F(\delta^*(w; d_c^*); D) \leq -[1 - h(D)] + \epsilon_3$ for all $w \in [\epsilon_1, \epsilon_2]$. Putting together this bound with the earlier bound (38) yields that for all $w \in [\epsilon_1, \epsilon_2]$:

$$\begin{aligned} K(w; d_c) &= R \frac{B(w)}{R_H} + F(\delta^*(w; d_c^*), D) \\ &\leq \mathrm{R}[1 - \rho(\epsilon_2)] - [1 - h(D)] + \epsilon_3 \\ &= \{\mathrm{R} - [1 - h(D)]\} + (\epsilon_3 - \mathrm{R}\rho(\epsilon_2)) \\ &= \Delta + (\epsilon_3 - \mathrm{R}\rho(\epsilon_2)) \end{aligned}$$

Since we are free to choose $\epsilon_3 > 0$, we may set $\epsilon_3 = \frac{\mathrm{R}\rho(\epsilon_2)}{2}$ to yield the claim.

6 Proof of channel coding optimality

In this section, we turn to the proof of the previously stated Theorem 2, concerning the channel coding optimality of the compound construction.

If the codeword $x \in \{0,1\}^n$ is transmitted, then the receiver observes $V = x \oplus W$, where W is a $\mathrm{Ber}(p)$ random vector. Our goal is to bound the probability that maximum likelihood (ML) decoding fails where the probability is taken over the randomness in both the channel noise and the code construction. To simplify the analysis, we focus on the following sub-optimal (non-ML) decoding procedure. Let ϵ_n be any non-negative sequence such that $\epsilon_n/n \to 0$ but $\epsilon_n^2/n \to +\infty$—say for instance, $\epsilon_n = n^{2/3}$.

Definition 2 (Decoding Rule:)**.** With the threshold $d(n) := pn + \epsilon_n$, decode to codeword $x_i \iff \|x_i \oplus V\|_1 \leq d(n)$, and no other codeword is within $d(n)$ of V.

The extra term ϵ_n in the threshold $d(n)$ is chosen for theoretical convenience. Using the following two lemmas, we establish that this procedure has arbitrarily small probability of error, whence ML decoding (which is at least as good) also has arbitrarily small error probability.

Lemma 7. *Using the suboptimal procedure specified in the definition* (2)*, the probability of decoding error vanishes asymptotically provided that*

$$R_G\, B(w) - D(p||\delta^*(w; d_c) * p) \;<\; 0 \qquad \textit{for all } w \in (0, \tfrac{1}{2}], \tag{39}$$

where B is any function bounding the average weight enumerator as in equation (34)*.*

Proof. Let $N = 2^{nR} = 2^{mR_H}$ denote the total number of codewords in the joint LDGM/LDPC code. Due to the linearity of the code construction and symmetry of the decoding procedure, we may assume without loss of generality that the all zeros codeword 0^n was transmitted (*i.e.*, $x = 0^n$). In this case, the channel output is simply $V = W$ and so our decoding procedure will fail if and only if one the following two conditions holds:

(i) either $\|W\|_1 > d(n)$, or

(ii) there exists a sequence of information bits $y \in \{0,1\}^m$ satisfying the parity check equation $Hy = 0$ such that the codeword Gy satisfies $\|Gy \oplus W\|_1 \leq d(n)$.

Consequently, using the union bound, we can upper bound the probability of error as follows:

$$p_{err} \leq \mathbb{P}[\|W\|_1 > d(n)] + \sum_{i=2}^{N} \mathbb{P}\big[\|Gy^i \oplus W\|_1 \leq d(n)\big]. \tag{40}$$

Since $\mathbb{E}[\|W\|_1] = pn$, we may apply Hoeffdings's inequality [13] to conclude that

$$\mathbb{P}[\|W\|_1 > d(n)] \quad \leq \quad 2\exp\left(-2\frac{\epsilon_n^2}{n}\right) \;\to\; 0 \tag{41}$$

by our choice of ϵ_n. Now focusing on the second term, let us rewrite it as a sum over the possible Hamming weights $\ell = 1, 2, \ldots, m$ of information sequences (i.e., $\|y\|_1 = \ell$) as follows:

$$\sum_{i=2}^{N} \mathbb{P}\big[\|Gy^i \oplus W\|_1 \leq d(n)\big] \quad = \quad \sum_{\ell=1}^{m} \mathbb{A}_m(\frac{\ell}{m})\, \mathbb{P}\big[\|Gy \oplus W\|_1 \geq d(n) \;\big|\; \|y\|_1 = \ell\big],$$

where we have used the fact that the (average) number of information sequences with fractional weight ℓ/m is given by the LDPC weight enumerator $\mathbb{A}_m(\frac{\ell}{m})$. Analyzing the probability terms in this sum, we note Lemma 9 (see Appendix A) guarantees that Gy has i.i.d. $\mathrm{Ber}(\delta^*(\frac{\ell}{m}; d_c))$ elements, where $\delta^*(\,\cdot\,; d_c)$ was defined in equation (29). Consequently, the vector $Gy \oplus W$ has i.i.d. $\mathrm{Ber}(\delta(\frac{\ell}{m}) * p)$ elements. Applying Sanov's theorem [11] for the special case of binomial variables yields that for any information bit sequence y with ℓ ones, we have

$$\mathbb{P}\big[\|Gy \oplus W\|_1 \geq d(n) \;\big|\; \|y\|_1 = \ell\big] \quad \leq \quad f(n) 2^{-nD\left(p||\delta(\frac{\ell}{m})*p\right)}, \tag{42}$$

for some polynomial term $f(n)$. We can then upper bound the second term in the error bound (40) as

$$\sum_{i=2}^{N} \mathbb{P}\big[\|Gy^i \oplus W\|_1 \leq d(n)\big] \quad \leq \quad f(m)\, \exp\left\{ \max_{1\leq\ell\leq m} \left[mB(\frac{\ell}{m}) + o(m) - nD\left(p||\delta(\frac{\ell}{m}) * p\right) \right] \right\},$$

where we have used equation (42), as well as the assumed upper bound (34) on $\mathbb{A}_m$ in terms of B. Simplifying further, we take logarithms and rescale by m to assess the exponential rate of decay,

thereby obtaining

$$\begin{aligned}\frac{1}{m}\log\sum_{i=2}^{N}\mathbb{P}\big[\|Gy^i\oplus W\|_1\le d(n)\big] &\le \max_{1\le\ell\le m}\Big[B(\frac{\ell}{m})-\frac{1}{R_G}D\left(p||\delta(\frac{\ell}{m})*p\right)\Big]+o(1)\\ &\le \max_{w\in[0,1]}\Big[B(w)-\frac{1}{R_G}D\left(p||\delta(w)*p\right)\Big]+o(1),\end{aligned}$$

and establishing the claim. □

Lemma 8. *For any $p\in(0,1)$ and total rate $R:=R_G\,R_H\ <\ 1-h\,(p)$, there exist finite choices of the degree triplet (d_c,d_v,d_c') such that (39) is satisfied.*

Proof. For notational convenience, we define

$$L(w)\ :=\ R_GB(w)-D\left(p||\delta^*(w;d_c)*p\right). \tag{43}$$

First of all, it is known [17] that a regular LDPC code with rate $R_H=\frac{d_v}{d_c'}<1$ and $d_v\ge 3$ has linear minimum distance. More specifically, there exists a threshold $\nu^*=\nu^*(d_v,d_c)$ such that $B(w)\le 0$ for all $w\in[0,\nu^*]$. Hence, since $B(w)-D\left(p||\delta^*(w;d_c)*p\right)\ge 0$ for all $w\in(0,1)$, for $w\in(0,\nu^*]$, we have $L(w)<0$.

Turning now to the interval $[\nu^*,\frac{1}{2}]$, consider the function

$$\widetilde{L}(w)\ :=\ Rh\,(w)-D\left(p||\delta^*(w;d_c)*p\right). \tag{44}$$

Since $B(w)\le R_Hh\,(w)$, we have $L(w)\le\widetilde{L}(w)$, so that it suffices to upper bound $\widetilde{L}$. Observe that $\widetilde{L}(\frac{1}{2})=R-(1-h\,(p))<0$ by assumption. Therefore, it suffices to show that, by appropriate choice of d_c, we can ensure that $\widetilde{L}(w)\le\widetilde{L}(\frac{1}{2})$. Noting that $\widetilde{L}$ is infinitely differentiable, calculating derivatives yields $\widetilde{L}'(\frac{1}{2})=0$ and $\widetilde{L}''(\frac{1}{2})<0$. (See Appendix G for details of these derivative calculations.) Hence, by second order Taylor series expansion around $w=\frac{1}{2}$, we obtain

$$\widetilde{L}(w)\ =\ \widetilde{L}(\frac{1}{2})+\frac{1}{2}\widetilde{L}''(\bar{w})(w-\frac{1}{2})^2,$$

where $\bar{w}\in[w,\frac{1}{2}]$. By continuity of $\widetilde{L}''$, we have $\widetilde{L}''(w)<0$ for all w in some neighborhood of $\frac{1}{2}$, so that the Taylor series expansion implies that $\widetilde{L}(w)\le\widetilde{L}(\frac{1}{2})$ for all w in some neighborhood, say $(\mu,\frac{1}{2}]$.

It remains to bound $\widetilde{L}$ on the interval $[\nu^*,\mu]$. On this interval, we have $\widetilde{L}(w)\le Rh\,(\mu)-D\left(p||\delta^*(\nu^*;d_c)*p\right)$. By examining equation (29) from Lemma 9, we see that by choosing d_c sufficiently large, we can make $\delta^*(\nu^*;d_c)$ arbitrarily close to $\frac{1}{2}$, and hence $D\left(p||\delta^*(\nu^*;d_c)*p\right)$ arbitrarily close to $1-h\,(p)$. More precisely, let us choose d_c large enough to guarantee that $D\left(p||\delta^*(\nu^*;d_c)*p\right)<(1-\epsilon)\,(1-h\,(p))$, where $\epsilon=\frac{R\,(1-h(\mu))}{1-h(p)}$. With this choice, we have, for all

$w \in [\nu^*, \mu]$, the sequence of inequalities

$$\begin{aligned}
\widetilde{L}(w) &\leq Rh(\mu) - D(p||\delta^*(\nu^*; d_c) * p) \\
&< Rh(\mu) - \left[(1 - h(p)) - R(1 - h(\mu))\right] \\
&= R - (1 - h(p)) < 0,
\end{aligned}$$

which completes the proof.

□

7 Discussion

In this paper, we established that it is possible to achieve both the rate-distortion bound for symmetric Bernoulli sources and the channel capacity for the binary symmetric channel using codes with bounded graphical complexity. More specifically, we have established that there exist low-density generator matrix (LDGM) codes and low-density parity check (LDPC) codes with finite degrees that, when suitably compounded to form a new code, are optimal for both source and channel coding. To the best of our knowledge, this is the first demonstration of classes of codes with bounded graphical complexity that are optimal as source and channel codes simultaneously. We also demonstrated that this compound construction has a naturally nested structure that can be exploited to achieve the Wyner-Ziv bound [45] for lossy compression of binary data with side information, as well as the Gelfand-Pinsker bound [19] for channel coding with side information.

Since the analysis of this paper assumed optimal decoding and encoding, the natural next step is the development and analysis of computationally efficient algorithms for encoding and decoding. Encouragingly, the bounded graphical complexity of our proposed codes ensures that they will, with high probability, have high girth and good expansion, thus rendering them well-suited to message-passing and other efficient decoding procedures. For pure channel coding, previous work [16, 36, 41] has analyzed the performance of belief propagation when applied to various types of compound codes, similar to those analyzed in this paper. On the other hand, for pure lossy source coding, our own past work [44] provides empirical demonstration of the feasibility of modified message-passing schemes for decoding of standard LDGM codes. It remains to extend both these techniques and their analysis to more general joint source/channel coding problems, and the compound constructions analyzed in this paper.

Acknowledgements

The work of MJW was supported by National Science Foundation grant CAREER-CCF-0545862, a grant from Microsoft Corporation, and an Alfred P. Sloan Foundation Fellowship.

A Basic property of LDGM codes

For a given weight $w \in (0,1)$, suppose that we enforce that the information sequence $y \in \{0,1\}^m$ has exactly $\lceil wm \rceil$ ones. Conditioned on this event, we can then consider the set of all codewords $X(w) \in \{0,1\}^n$, where we randomize over low-density generator matrices G chosen as in step (a) above. Note for any fixed code, $X(w)$ is simply some codeword, but becomes a random variable when we imagine choosing the generator matrix G randomly. The following lemma characterizes this distribution as a function of the weight w and the LDGM top degree d_c:

Lemma 9. *Given a binary vector $y \in \{0,1\}^m$ with a fraction w of ones, the distribution of the random LDGM codeword $X(w)$ induced by y is i.i.d. Bernoulli with parameter $\delta^*(w; d_c) = \frac{1}{2}\left[1 - (1-2w)^{d_c}\right]$.*

Proof. Given a fixed sequence $y \in \{0,1\}^m$ with a fraction w ones, the random codeword bit $X_i(w)$ at bit i is formed by connecting d_c edges to the set of information bits.[3] Each edge acts as an i.i.d. Bernoulli variable with parameter w, so that we can write

$$X_i(w) \quad = \quad V_1 \oplus V_2 \oplus \ldots \oplus V_{d_c}, \tag{45}$$

where each $V_k \sim \mathrm{Ber}(w)$ is independent and identically distributed. A straightforward calculation using z-transforms (see [17]) or Fourier transforms over $GF(2)$ yields that $X_i(w)$ is Bernoulli with parameter $\delta^*(w; d_c)$ as defined. □

B Bounds on binomial coefficients

The following bounds on binomial coefficients are standard (see Chap. 12, [11]):

$$h\left(\frac{k}{n}\right) - \frac{\log(n+1)}{n} \;\leq\; \frac{1}{n}\log\binom{n}{k} \;\leq\; h\left(\frac{k}{n}\right). \tag{46}$$

Here, for $\alpha \in (0,1)$, the quantity $h(\alpha) := -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$ is the binomial entropy function.

[3] In principle, our procedure allows two different edges to choose the same information bit, but the probability of such double-edges is asymptotically negligible.

C Proof of Lemma 4

First, by the definition of $T_n(D)$, we have

$$\begin{aligned}\mathbb{E}[T_n^2(D)] &= \mathbb{E}\left[\sum_{i=1}^{N-1}\sum_{j=0}^{N-1} Z^i(D)Z^j(D)\right] \\ &= \mathbb{E}[T_n] + \sum_{i=0}^{N-1}\sum_{j\neq i}\mathbb{P}[Z^i(D)=1,\, Z^i(D)=1].\end{aligned}$$

To simplify the second term on the RHS, we first note that for any i.i.d Bernoulli($\frac{1}{2}$) sequence $S \in \{0,1\}^n$ and any codeword X^j, the binary sequence $S' := S \oplus X^j$ is also i.i.d. Bernoulli($\frac{1}{2}$). Consequently, for each pair $i \neq j$, we have

$$\begin{aligned}\mathbb{P}\left[Z^i(D)=1,\, Z^j(D)=1\right] &= \mathbb{P}\left[\|X^i \oplus S\|_1 \leq Dn, \|X^j \oplus S\|_1 \leq Dn\right] \\ &= \mathbb{P}\left[\|X^i \oplus S'\|_1 \leq Dn, \|X^j \oplus S'\|_1 \leq Dn\right] \\ &= \mathbb{P}\left[\|X^i \oplus X^j \oplus S\|_1 \leq Dn, \|S\|_1 \leq Dn\right].\end{aligned}$$

Note that for each $j \neq i$, the vector $X^i \oplus X^j$ is a non-zero codeword. For each fixed i, summing over $j \neq i$ can be recast as summing over all non-zero codewords, so that

$$\begin{aligned}\sum_{i\neq j}\mathbb{P}\left[Z^i(D)=1,\, Z^j(D)=1\right] &= \sum_{i=0}^{N-1}\sum_{j\neq i}\mathbb{P}\left[\|X^i \oplus X^j \oplus S\|_1 \leq Dn, \|S\|_1 \leq Dn\right] \\ &= \sum_{i=0}^{N-1}\sum_{k\neq 0}\mathbb{P}\left[\|X^k \oplus S\|_1 \leq Dn, \|S\|_1 \leq Dn\right] \\ &= 2^{n\mathrm{R}}\sum_{k\neq 0}\mathbb{P}\left[\|X^k \oplus S\|_1 \leq Dn, \|S\|_1 \leq Dn\right] \\ &= 2^{n\mathrm{R}}\mathbb{P}\left[Z^0(D)=1\right]\sum_{k\neq 0}\mathbb{P}\left[Z^k(D)=1 \mid Z^0(D)=1\right] \\ &= \mathbb{E}[T_n]\sum_{k\neq 0}\mathbb{P}\left[Z^k(D)=1 \mid Z^0(D)\right]\end{aligned}$$

thus establishing the claim.

D Proof of Lemma 5

We reformulate the probability $\mathbb{Q}(w, D)$ as follows. Recall that $\mathbb{Q}$ involves conditioning the source sequence S on the event $\|S\|_1 \leq Dn$. Accordingly, we define a discrete variable T with distribution

$$\mathbb{P}(T = t) \quad = \quad \frac{\binom{n}{t}}{\sum_{s=0}^{Dn} \binom{n}{s}} \qquad \text{for } t = 0, 1, \ldots, Dn,$$

representing the (random) number of 1s in the source sequence S. Let U_i and V_j denote Bernoulli random variables with parameters $1 - \delta^*(w; d_c)$ and $\delta^*(w; d_c)$ respectively. With this set-up, conditioned on codeword j having a fraction wn ones, the quantity $\mathbb{Q}(w, D)$ is equivalent to the probability that the random variable

$$W \quad := \quad \begin{cases} \sum_{i=1}^{T} U_j + \sum_{j=1}^{n-T} V_j & \text{if } T \geq 1 \\ \sum_{j=1}^{n} V_j & \text{if } T = 0 \end{cases} \tag{47}$$

is less than Dn. To bound this probability, we use a Chernoff bound in the form

$$\frac{1}{n} \log \mathbb{P}[W \leq Dn] \quad \leq \quad \inf_{\lambda < 0} \left(\frac{1}{n} \log \mathbb{M}_W(\lambda) - \lambda D \right). \tag{48}$$

We begin by computing the moment generating function $\mathbb{M}_W$. Taking conditional expectations and using independence, we have

$$\mathbb{M}_W(\lambda) \quad = \quad \sum_{t=0}^{Dn} \mathbb{P}[T = t] \; [\mathbb{M}_U(\lambda)]^t \, [\mathbb{M}_V(\lambda)]^{n-t}.$$

Here the cumulant generating functions have the form

$$\log \mathbb{M}_U(\lambda) \quad = \quad \log \left[(1 - \delta) e^{\lambda} + \delta \right], \quad \text{and} \tag{49a}$$

$$\log \mathbb{M}_V(\lambda) \quad = \quad \log \left[(1 - \delta) + \delta e^{\lambda} \right], \tag{49b}$$

where we have used (and will continue to use) δ as a shorthand for $\delta^*(w; d_c)$.

Of interest to us is the exponential behavior of this expression in n. Using the standard entropy approximations to the binomial coefficient (see Appendix B), we can bound $\mathbb{M}_W(\lambda)$ as

$$f(n) \sum_{t=0}^{Dn} \underbrace{\exp \left[n \left\{ h\left(\frac{t}{n}\right) - h\left(D\right) + \frac{t}{n} \log \mathbb{M}_U(\lambda) + \left(1 - \frac{t}{n}\right) \log \mathbb{M}_V(\lambda) \right\} \right]}_{g(t)}, \tag{50}$$

where $f(n)$ denotes a generic polynomial factor. Further analyzing this sum, we have

$$\begin{aligned}\frac{1}{n}\log\sum_{t=0}^{Dn} g(t) &\leq \frac{1}{n}\max_{0\leq t\leq Dn}\log g(t) + \frac{\log f(n)}{n} + \frac{\log(nD)}{n} \\ &= \max_{0\leq t\leq Dn}\left\{h\left(\frac{t}{n}\right) - h\left(D\right) + \frac{t}{n}\log\mathbb{M}_U(\lambda) + \left(1-\frac{t}{n}\right)\log\mathbb{M}_V(\lambda)\right\} + o(1) \\ &\leq \max_{u\in[0,D]}\left\{h\left(u\right) - h\left(D\right) + u\log\mathbb{M}_U(\lambda) + (1-u)\log\mathbb{M}_V(\lambda)\right\} + o(1).\end{aligned}$$

Combining this upper bound on $\frac{1}{n}\log\mathbb{M}_W(\lambda)$ with the Chernoff bound (48) yields that

$$\frac{1}{n}\log\mathbb{P}[W\leq Dn] \leq \inf_{\lambda<0}\max_{u\in[0,D]} G(u,\lambda;\delta) + o(1) \tag{51}$$

where the function G takes the form

$$G(u,\lambda;\delta) := h\left(u\right) - h\left(D\right) + u\log\mathbb{M}_U(\lambda) + (1-u)\log\mathbb{M}_V(\lambda) - \lambda D. \tag{52}$$

Finally, we establish that the solution (u^*,λ^*) to the min-max saddle point problem (51) is unique, and specified by $u^* = D$ and λ^* as in Lemma 5. First of all, observe that for any $\delta\in(0,1)$, the function G is continuous, strictly concave in u and strictly convex in λ. (The strict concavity follows since $h\left(u\right)$ is strictly concave with the remaining terms linear; the strict convexity follows since cumulant generating functions are strictly convex.) Therefore, for any fixed $\lambda<0$, the maximum over $u\in[0,D]$ is always achieved. On the other hand, for any $D>0$, $u\in[0,D]$ and $\delta\in(0,1)$, we have $G(u;\lambda;t)\to+\infty$ as $\lambda\to-\infty$, so that the infimum is either achieved at some $\lambda^*<0$, or at $\lambda^*=0$. We show below that it is always achieved at an interior point $\lambda^*<0$. Thus far, using standard saddle point theory [21], we have established the existence and uniqueness of the saddle point solution (u^*,λ^*).

To verify the fixed point conditions, we compute partial derivatives in order to find the optimum. First, considering u, we compute

$$\begin{aligned}\frac{\partial G}{\partial u}(u,\lambda;\delta) &= \log\frac{1-u}{u} + \log\mathbb{M}_U(\lambda) - \log\mathbb{M}_V(\lambda) \\ &= \log\frac{1-u}{u} + \log\left[(1-\delta)e^{\lambda}+\delta\right] - \log\left[(1-\delta)+\delta e^{\lambda}\right].\end{aligned}$$

Solving the equation $\frac{\partial G}{\partial u}(u,\lambda;\delta)=0$ yields

$$u' = \frac{\exp(\lambda)}{1+\exp(\lambda)}D + \frac{1}{1+\exp(\lambda)}(1-D) \geq 0. \tag{53}$$

Since $D\leq\frac{1}{2}$, a bit of algebra shows that $u'\geq D$ for all choices of λ. Since the maximization is

constrained to $[0, D]$, the optimum is always attained at $u^* = D$.

Turning now to the minimization over λ, we compute the partial derivative to find

$$\frac{\partial G}{\partial \lambda}(u, \lambda; \delta) \quad = \quad u\frac{(1-\delta)\exp(\lambda)}{(1-\delta)\exp(\lambda)+\delta} + (1-u)\frac{\delta\exp(\lambda)}{(1-\delta)+\delta\exp(\lambda)} - D.$$

Setting this partial derivative to zero yields a quadratic equation in $\exp(\lambda)$ with coefficients

$$a \quad = \quad \delta\,(1-\delta)\,(1-D) \tag{54a}$$

$$b \quad = \quad u(1-\delta)^2 + (1-u)\delta^2 - D\left[\delta^2 + (1-\delta)^2\right]. \tag{54b}$$

$$c \quad = \quad -D\delta(1-\delta). \tag{54c}$$

The unique positive root ρ^* of this quadratic equation is given by

$$\rho^*(\delta, D, u) \quad := \quad \frac{1}{2a}\left[-b + \sqrt{b^2 - 4ac}\right]. \tag{55}$$

It remains to show that $\rho^* \leq 1$, so that $\lambda^* := \log \rho^* < 0$. A bit of algebra (using the fact $a \geq 0$) shows that $\rho^* < 1$ if and only if $a + b + c > 0$. We then note that at the optimal $u^* = D$, we have $b = (1-2D)\delta^2$, whence

$$\begin{aligned} a + b + c \quad &= \quad \delta\,(1-\delta)\,(1-D) + (1-2D)\delta^2 - D\delta(1-\delta) \\ &= \quad (1-2D)\,\delta \;>\; 0 \end{aligned}$$

since $D < \frac{1}{2}$ and $\delta > 0$. Hence, the optimal solution is $\lambda^* := \log \rho^* < 0$, as specified in the lemma statement.

E Proof of Lemma 6

A straightforward calculation yields that

$$G(\frac{1}{2}) = F(\delta^*(\frac{1}{2}; d_c); D) = F(\frac{1}{2}; D) = -\left(1 - h\left(D\right)\right)$$

as claimed. Turning next to the derivatives, we note that by inspection, the solution $\lambda^*(t)$ defined in Lemma 5 is twice continuously differentiable as a function of t. Consequently, the function $F(t, D)$ is twice continuously differentiable in t. Moreover, the function $\delta^*(w; d_c)$ is twice continuously differentiable in w. Overall, we conclude that $G(w) = F(\delta^*(w; d_c); D)$ is twice continuously differentiable in w, and that we can obtain derivatives via chain rule. Computing the first derivative,

we have

$$G'(\frac{1}{2}) \;\; = \;\; \delta'(\frac{1}{2})\, F'(\delta^*(\frac{1}{2}; d_c); D) \;\; = \;\; 0$$

since $\delta'(w) = -d_c\,(1-2w)^{d_c-1}$, which reduces to zero at $w = \frac{1}{2}$. Turning to the second derivative, we have

$$G''(\frac{1}{2}) \;\; = \;\; \delta''(\frac{1}{2})\, F'(\delta^*(\frac{1}{2}; d_c); D) + \left(\delta'(\frac{1}{2})\right)^2 F''(\delta^*(\frac{1}{2}; d_c); D) \;\; = \;\; \delta''(\frac{1}{2})\, F'(\delta^*(\frac{1}{2}; d_c); D).$$

We again compute $\delta''(w) = 2d_c\,(d_c-1)(1-2w)^{d_c-2}$, which again reduces to zero at $w = \frac{1}{2}$ since $d_c \geq 4$ by assumption.

F Regular LDPC codes are sufficient

Consider a regular (d_v, d_c') code from the standard Gallager LDPC ensemble. In order to complete the proof of Theorem 1, we need to show for suitable choices of degree (d_v, d_c'), the average weight enumerator of these codes can be suitably bounded, as in equation (34), by a function B that satisfies the conditions specified in Section 5.5.

It can be shown [17, 22] that for even degrees d_c', the average weight enumerator of the regular Gallager ensemble, for any block length m, satisfies the bound

$$\frac{1}{m} \log \mathbb{A}_m(w) \;\; = \;\; B(w; d_v, d_c') + o(1).$$

The function B in this relation is defined for $w \in [0, \frac{1}{2}]$ as

$$B(w; d_v, d_c') \;\; := \;\; (1-d_v)h\,(w) - (1-R_H) + d_v \inf_{\lambda \leq 0} \left\{ \frac{1}{d_c'} \log \left((1+e^{\lambda})^{d_c'} + (1-e^{\lambda})^{d_c'} \right) - w\lambda \right\}, \quad (56)$$

and by $B(w) = B(w - \frac{1}{2})$ for $w \in [\frac{1}{2}, 1]$. Given that the minimization problem (56) is strictly convex, a straightforward calculation of the derivative shows the optimum is achieved at λ^*, where $\lambda^* \leq 0$ is the unique solution of the equation

$$e^{\lambda} \frac{(1+e^{\lambda})^{d_c'-1} - (1-e^{\lambda})^{d_c'-1}}{(1+e^{\lambda})^{d_c'} + (1-e^{\lambda})^{d_c'}} \;\; = \;\; w. \quad (57)$$

Some numerical computation for $R_H = 0.5$ and different choices (d_v, d_c') yields the curves shown in Fig. 9.

We now show that for suitable choices of degree (d_v, d_c'), the function B defined in equation (56) satisfies the four assumptions specified in Section 5.5. First, for even degrees d_c', the function B

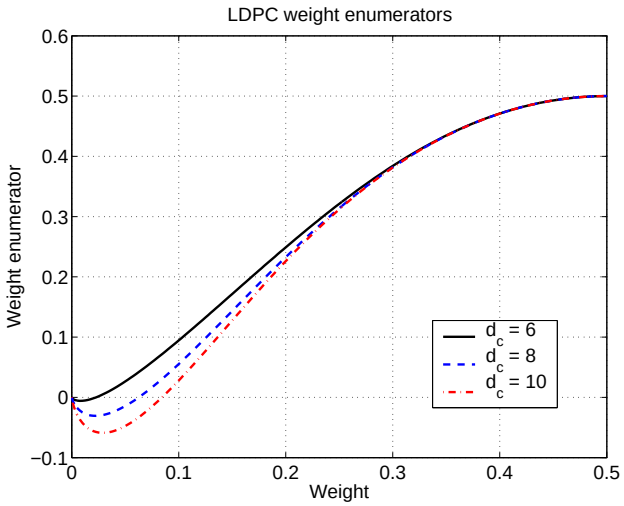


Figure 9. Plots of LDPC weight enumerators for codes of rate $R_H = 0.5$, and check degrees $d_c' \in \{6, 8, 10\}$.

is symmetric about $w = \frac{1}{2}$, so that assumption (A1) holds. Secondly, we have $B(w) \leq R_H$, and moreover, for $w = \frac{1}{2}$, the optimal $\lambda^*(\frac{1}{2}) = 0$, so that $B(\frac{1}{2}) = R_H$, and assumption (A3) is satisfied. Next, it is known from the work of Gallager [17], and moreover is clear from the plots in Fig. 9, that LDPC codes with $d_v > 2$ have linear minimum distance, so that assumption (A4) holds.

The final condition to verify is assumption (A2), concerning the differentiability of B. We summarize this claim in the following:

Lemma 10. *The function B is twice continuously differentiable on $(0, 1)$, and in particular we have*

$$B'(\frac{1}{2}) \;=\; 0, \qquad \textit{and} \qquad B''(\frac{1}{2}) \;<\; 0. \tag{58}$$

Proof. Note that for each fixed $w \in (0, 1)$, the function

$$f(\lambda) \;\;=\;\; \frac{1}{d_c'} \log \left((1 + e^{\lambda})^{d_c'} + (1 - e^{\lambda})^{d_c'} \right) \;\;=\;\; \frac{1}{d_c'} \log \left((e^{-\lambda} + 1)^{d_c'} + (e^{-\lambda} - 1)^{d_c'} \right) + \lambda$$

is strictly convex and twice continuously differentiable as a function of λ. Moreover, the function $f^*(w) := \inf_{\lambda \leq 0} \{f(\lambda) - \lambda w\}$ corresponds to the conjugate dual [21] of $f(\lambda) + \mathbb{I}_{\leq 0}(\lambda)$. Since the optimum is uniquely attained for each $w \in (0, 1)$, an application of Danskin's theorem [4] yields that f^* is differentiable with $\frac{d}{dw} f^*(w) = -\lambda^*(w)$, where λ^* is defined by equation (57). Putting together the pieces, we have $B'(w) = (1 - d_v)h'(w) - d_v \lambda^*(w)$. Evaluating at $w = \frac{1}{2}$ yields $B'(\frac{1}{2}) = 0 - d_v \lambda^*(0) = 0$ as claimed.

We now claim that $\lambda^*(w)$ is differentiable. Indeed, let us write the defining relation (57) for $\lambda^*(w)$ as $F(\lambda, w) = 0$ where $F(\lambda, w) := f'(\lambda) - w$. Note that F is twice continuously differentiable

in both λ and w; moreover, $\frac{\partial F}{\partial \lambda}$ exists for all $\lambda \leq 0$ and w, and satisfies $\frac{\partial F}{\partial \lambda}(\lambda, w) = f''(\lambda) > 0$ by the strict convexity of f. Hence, applying the implicit function theorem [4] yields that $\lambda^*(w)$ is differentiable, and moreover that $\frac{d\lambda^*}{dw}(w) = 1/f''(\lambda^*(w))$. Hence, combined with our earlier calculation of B', we conclude that $B''(w) = (1 - d_v)h''(w) - d_v \frac{1}{f''(\lambda(w))}$. Our final step is to compute the second derivative f''. In order to do so, it is convenient to define $g = \log f'$, and exploit the relation $g'f' = f''$. By definition, we have

$$g(\lambda) \;=\; \lambda + \log\left[(1+e^{\lambda})^{d'_c - 1} - (1-e^{\lambda})^{d'_c - 1}\right] - \log\left[(1+e^{\lambda})^{d'_c} + (1-e^{\lambda})^{d'_c}\right]$$

whence

$$g'(\lambda) \;=\; 1 + e^{\lambda}(d'_c - 1)\frac{(1+e^{\lambda})^{d'_c - 2} + (1-e^{\lambda})^{d'_c - 2}}{(1+e^{\lambda})^{d'_c - 1} - (1-e^{\lambda})^{d'_c - 1}} - e^{\lambda} d'_c \frac{(1+e^{\lambda})^{d'_c - 1} - (1-e^{\lambda})^{d'_c - 1}}{(1+e^{\lambda})^{d'_c} + (1-e^{\lambda})^{d'_c}}$$

Evaluating at $w = \frac{1}{2}$ corresponds to $\lambda(0) = 0$, so that

$$f''(\lambda(\frac{1}{2})) \;=\; f'(0)\, g'(0) \;=\; \frac{1}{2}\left[1 + (d'_c - 1)\frac{2^{d'_c - 2}}{2^{d'_c - 1}} - d'_c \frac{2^{d'_c - 1}}{2^{d'_c}}\right] \;=\; \frac{1}{4}.$$

Consequently, combining all of the pieces, we have

$$B''(w) = (1 - d_v)h''(\frac{1}{2}) - d_v \frac{1}{f''(\lambda(\frac{1}{2}))} \;=\; \frac{d_v - 1}{4} - 4d_v < 0$$

as claimed.

□

G Derivatives of $\widetilde{L}$

Here we calculate the first and second derivatives of the function $\widetilde{L}$ defined in equation (44). The first derivative takes the form

$$\widetilde{L}'(v) \;=\; R \log \frac{1-v}{v} + p\frac{\delta'(v; d_c)}{\delta(v; d_c)} - (1-p)\frac{\delta'(v; d_c)}{1 - \delta(v; d_c)}$$

where $\delta'(v; d_c) = d_c(1-2v)^{d_c-1}$. Since $\delta'(\frac{1}{2}; d_c) = 0$, we have $\widetilde{L}'(\frac{1}{2}) = 0$ as claimed. Second, using chain rule, we calculate

$$\widetilde{L}''(v) = -R\big[\frac{1}{1-v} + \frac{1}{v}\big] + p\frac{\delta''(v;d_c)\delta(v;d_c) - [\delta'(v;d_c)]^2}{[\delta(v;d_c)]^2} - (1-p)\frac{\delta''(v;d_c)\big[1-\delta(v;d_c)\big] + [\delta'(v;d_c)]^2}{[1-\delta(v;d_c)]^2}$$

and $\delta''(v; d_c) = -d_c\,(d_c-1)\,(1-2v)^{d_c-2}$. Now for $d_c > 2$, we have $\delta''(\frac{1}{2}) = 0$, so that $\widetilde{L}''(\frac{1}{2}) = -4R < 0$ as claimed.

References

[1] N. Alon and J. Spencer. *The Probabilistic Method*. Wiley Interscience, New York, 2000.

[2] R. J. Barron, B. Chen, and G. W. Wornell. The duality between information embedding and source coding with side information and some applications. *IEEE Trans. Info. Theory*, 49(5):1159–1180, 2003.

[3] C. Berroux and A. Glavieux. Near optimum error correcting coding and decoding: Turbo codes. *IEEE Trans. Commun.*, 44:1261–1271, October 1996.

[4] D. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

[5] J. Chou, S. S. Pradhan, and K. Ramchandran. Turbo coded trellis-based constructions for data embedding: Channel coding with side information. In *Proceedings of the Asilomar Conference*, November 2001.

[6] J. Chou, S. S. Pradhan, and K. Ramchandran. Turbo and trellis-based constructions for source coding with side information. In *Proceedings of the Data Compression Conference (DCC)*, 2003.

[7] S.-Y. Chung, G. D. Forney, T. Richardson, and R. Urbanke. On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit. *IEEE Communications Letters*, 5(2):58–60, February 2001.

[8] S. Ciliberti and M. Mézard. The theoretical capacity of the parity source coder. Technical report, August 2005. arXiv:cond-mat/0506652.

[9] S. Ciliberti, M. Mézard, and R. Zecchina. Message-passing algorithms for non-linear nodes and data compression. Technical report, November 2005. arXiv:cond-mat/0508723.

[10] S. Cocco, O. Dubois, J. Mandler, and R. Monasson. Rigorous decimation-based construction of ground pure states for spin-glass models on random lattices. *Physical Review Letters*, 90(4), January 2003.

[11] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[12] N. Creignou, H. Daud/'e, and O. Dubois. Approximating the satisfiability threshold of random XOR formulas. *Combinatorics, Probability and Computing*, 12:113–126, 2003.

[13] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[14] O. Dubois and J. Mandler. The 3-XORSAT threshold. In *Proc. 43rd Symp. FOCS*, pages 769–778, 2002.

[15] U. Erez and S. ten Brink. A close-to-capacity dirty paper coding scheme. *IEEE Trans. Info. Theory*, 51(10):3417–3432, 2005.

[16] O. Etesami and A. Shokrollahi. Raptor codes on binary memoryless symmetric channels. *IEEE Trans. on Information Theory*, 52(5):2033–2051, 2006.

[17] R. G. Gallager. *Low-density parity check codes*. MIT Press, Cambridge, MA, 1963.

[18] J. Garcia-Frias and Y. Zhao. Compression of binary memoryless sources using punctured turbo codes. *IEEE Communication Letters*, 6(9):394–396, September 2002.

[19] S. I. Gelfand and M. S. Pinsker. Coding for channel with random parameters. *Probl. Pered. Inform. (Probl. Inf. Tranmission)*, 9(1):19–31, 1983.

[20] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford Science Publications, Clarendon Press, Oxford, 1992.

[21] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume 1. Springer-Verlag, New York, 1993.

[22] S. Litsyn and V. Shevelev. On ensembles of low-density parity-check codes: asymptotic distance distributions. *IEEE Trans. Info. Theory*, 48(4):887–908, April 2002.

[23] A. Liveris, Z. Xiong, and C. Georghiades. Nested convolutional/turbo codes for the binary Wyner-Ziv problem. In *Proceedings of the International Conference on Image Processing (ICIP)*, volume 1, pages 601–604, September 2003.

[24] H. A. Loeliger. An introduction to factor graphs. *IEEE Signal Processing Magazine*, 21:28–41, 2004.

[25] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. Spielman. Improved low-density parity check codes using irregular graphs. *IEEE Trans. Info. Theory*, 47:585–598, February 2001.

[26] M. W. Marcellin and T. R. Fischer. Trellis coded quantization of memoryless and Gauss-Markov sources. *IEEE Trans. Communications*, 38(1):82–93, 1990.

[27] E. Martinian and M. J. Wainwright. Analysis of LDGM and compound codes for lossy compression and binning. In *Workshop on Information Theory and Applications (ITA)*, February 2006. Available at arxiv:cs.IT/0602046.

[28] E. Martinian and M. J. Wainwright. Low density codes achieve the rate-distortion bound. In *Data Compression Conference*, volume 1, March 2006. Available at arxiv:cs.IT/061123.

[29] E. Martinian and M. J. Wainwright. Low density codes can achieve the Wyner-Ziv and Gelfand-Pinsker bounds. In *International Symposium on Information Theory*, July 2006. Available at arxiv:cs.IT/0605091.

[30] E. Martinian and J. Yedidia. Iterative quantization using codes on graphs. In *Allerton Conference on Control, Computing, and Communication*, October 2003.

[31] Y. Matsunaga and H. Yamamoto. A coding theorem for lossy data compression by LDPC codes. *IEEE Trans. Info. Theory*, 49:2225–2229, 2003.

[32] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina. Alternative solutions to diluted p-spin models and XORSAT problems. *Jour. of Statistical Physics*, 111:105, 2002.

[33] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[34] T. Murayama. Thouless-Anderson-Palmer approach for lossy compression. *Physical Review E*, 69:035105(1)–035105(4), 2004.

[35] T. Murayama and M. Okada. One step RSB scheme for the rate distortion function. *J. Phys. A: Math. Gen.*, 65:11123–11130, 2003.

[36] H. Pfister, I. Sason, and R. Urbanke. Capacity-achieving ensembles for the binary erasure channel with bounded complexity. *IEEE Trans. on Information Theory*, 51(7):2352–2379, 2005.

[37] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): Design and construction. *IEEE Trans. Info. Theory*, 49(3):626–643, 2003.

[38] T. Richardson, A. Shokrollahi, and R. Urbanke. Design of capacity-approaching irregular low-density parity check codes. *IEEE Trans. Info. Theory*, 47:619–637, February 2001.

[39] T. Richardson and R. Urbanke. The capacity of low-density parity check codes under message-passing decoding. *IEEE Trans. Info. Theory*, 47:599–618, February 2001.

[40] D. Schonberg, S. S. Pradhan, and K. Ramchandran. LDPC codes can approach the slepian-wolf bound for general binary sources. In *Proceedings of the 40th Annual Allerton Conference on Control, Communication, and Computing*, pages 576–585, October 2002.

[41] A. Shokrollahi. Raptor codes. *IEEE Trans. on Information Theory*, 52(6):2551–2567, 2006.

[42] Y. Sun, A. Liveris, V. Stankovic, and Z. Xiong. Near-capacity dirty-paper code designs based on TCQ and IRA codes. In *ISIT*, September 2005.

[43] A. J. Viterbi and J. K. Omura. Trellis encoding of memoryless discrete-time sources with a fidelity criterion. *IEEE Trans. Info. Theory*, IT-20(3):325–332, 1974.

[44] M. J. Wainwright and E. Maneva. Lossy source coding by message-passing and decimation over generalized codewords of LDGM codes. In *International Symposium on Information Theory*, Adelaide, Australia, September 2005. Available at arxiv:cs.IT/0508068.

[45] A. D. Wyner and J. Ziv. The rate-distortion function for source encoding with side information at the encoder. *IEEE Trans. Info. Theory*, IT-22:1–10, January 1976.

[46] Y. Yang, V. Stankovic, Z. Xiong, and W. Zhao. On multiterminal source code design. In *Proceedings of the Data Compression Conference*, 2005.

[47] R. Zamir, S. S. (Shitz), and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Trans. Info. Theory*, 6(48):1250–1276, 2002.